DOCUMENT RESUME

ED 044 371 SP 004 390

AUTHOR	Morsh, Joseph E.; Wilder, Eleanor W.
TITLE	Identifying the Effective Instructor: A Review of the Quantitative Studies. 1900-1952.
INSTITUTION	Air Force Personnel and Training Research Center, Chanute AFB, Ill.
REPORT NO	AFPTRC-TR-54-44
PUB DATE	54
NOTE	159p.
EDRS PRICE	EDRS Price MF-$0.75 HC-$8.05
DESCRIPTORS	*Educational Research, *Effective Teaching, *Evaluation Criteria, *Teacher Characteristics, *Teacher Evaluation

ABSTRACT

This research review contains summary and synthesis of 360 references selected from over 900 in 1) Education Index, 2) Psychological Abstracts, and 3) some 40 reviews and bibliographies, 28 of which were selected for inclusion in the 392-item bibliography at the end of this review. Principal findings of the cited quantitative research studies are summarized in the introductory section. Concluding implications for further research, presented as a guide in Air Force technical training research projects, are also expected to assist other investigators in the field. The description of research studies and tabular material are presented chronologically (1900-1952) under each topic heading. Topics under the major heading of "Criteria for Instructor Effectiveness" are rating the effectiveness of instructors, administrator rating, peer rating, student rating, self-rating, objective observation of performance, and student change as a measure. Topics under "The Predictors--Traits and Qualities Assumed to be Related to Instructor Effectiveness" are intelligence, education, scholarship, age and experience, knowledge of subject matter and present professional information and teacher examination scores, extracurricular activities and general culture test scores, socioeconomic status and sex and marital status, teaching aptitude and attitude toward teaching and interest, voice and speech characteristics, photograph, statistical analyses of abilities, personality studies and tests. (JS)

OCT 12 1970

ED0 44371

RESEARCH

Bulletin

AFPTRC-TR-54-44

Identifying the Effective Instructor: A Review Of the Quantitative Studies, 1900-1952

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

By Joseph E. Morsh
And Eleanor W. Wilder

Reprinted June 1967

AIR FORCE PERSONNEL & TRAINING RESEARCH CENTER
LACKLAND AIR FORCE BASE • SAN ANTONIO • TEXAS

SP004390

HEADQUARTERS
AIR FORCE PERSONNEL AND TRAINING RESEARCH CENTER
Air Research and Development Command
Lackland Air Force Base, Texas

Col. Herbert N. Cowles
Commander

Dr. Arthur W. Melton
Technical Director

Dr. Charles W. Bray
Deputy
For Research and Development

Dr. Henry S. Odbert
Director
For Technical Information

RESEARCH Bulletin ▲ AFPTRC-TR-54-44 OCTOBER 1954

The armed services now require that large numbers of personnel be rapidly trained to a high degree of technical competence in the operation and maintenance of a variety of complex electronic and mechanical equipment. The rapidity and effectiveness of training rest in large measure upon the quality of instruction. Hence considerable importance is assumed by the problems of locating individuals among military personnel who may be trained to become effective instructors and of evaluating instructors on the job. As a first step, the criteria suitable for determining instructor proficiency must be ascertained, and the factors employed to predict instructor effectiveness evaluated. In this Research Bulletin over three hundred sixty civilian studies which relate to the evaluation and prediction of instructor proficiency are reviewed and interpreted.

Dr. Joseph E. Morsh is a member of the Personnel Research Laboratory of this Center. Mrs. Eleanor W. Wilder was formerly a member of the Training Aids Research Laboratory. The survey of studies was made as part of the regular research of the Training Aids Research Laboratory.

ED044371

IDENTIFYING THE EFFECTIVE INSTRUCTOR:
A REVIEW OF THE QUANTITATIVE STUDIES
1900-1952

By Joseph E. Morsh
And Eleanor W. Wilder

Training Aids Research Laboratory
AIR FORCE PERSONNEL AND TRAINING RESEARCH CENTER
Air Research and Development Command
Chanute Air Force Base, Illinois

Approved by:
Richard W. Faubion, Col, USAF, Director
Arthur A. Lumsdaine, Technical Director
Training Aids Research Laboratory

Project No. 7714
Task No. 77243

ACKNOWLEDGMENTS

This review is the outgrowth of a working bibliography which was assembled in connection with Human Resources Research Center Project 507-010-0005, "The Identification of the Characteristics and Teaching Procedures of Successful Technical School Instructor." The authors are grateful to Dr. George B. Simon and to Dr. Robert A. Swanson who read the manuscript and offered many constructive criticisms. Special acknowledgment is due to Mr. James R. Berkshire who worked closely with the senior author in the preparation of the manuscript. Mrs. Katherine M. Zawadke supervised the setting up of the tables.

TABLE OF CONTENTS

Page

List of Tables iv

Introduction 1

Principal Findings of Cited Research Studies 2

- Criteria 2
- Predictors 5

Criteria of Instructor Effectiveness 7

- Rating the Effectiveness of Instructors 10
- Administrative Rating of Instructor Effectiveness 15
- Peer Rating of Instructor Effectiveness 23
- Student Rating of Instructor Effectiveness 27
- Self-Rating of Instructor Effectiveness 40
- Objective Observation of Instructor Performance 42
- Student Change as a Measure of Instructor Effectiveness 50

The Predictors--Traits and Qualities Assumed to be Related to Instructor Effectiveness 59

- Intelligence as Related to Instructor Effectiveness 60
- Education as Related to Instructor Effectiveness 66
- Scholarship as Related to Instructor Effectiveness 70
- Age and Experience as Related to Instructor Effectiveness 79
- Knowledge of Subject Matter, Present Professional Information, and Teacher Examination Scores as Related to Instructor Effectiveness 84
- Extracurricular Activities and General Culture Test Scores versus Instructor Effectiveness 87
- Socioeconomic Status, Sex, and Marital Status versus Instructor Effectiveness 91
- The Relation of Teaching Aptitude, Attitude Toward Teaching, and Interest to Instructor Effectiveness 96
- The Relation of Voice and Speech Characteristics to Instructor Effectiveness 101
- The Photograph as a Predictor of Instructor Effectiveness 104
- Statistical Analyses of Instructor Abilities 105
- Opinion Studies of the Personality Characteristics of Effective and Ineffective Instructors 108
- Personality Tests of Teachers 114

Implications for Further Research 118

Table of Contents (Cont.)

Page

Criterion Research 119
Predictor Research 122

Bibliography 125

Reviews and Bibliographies 149

LIST OF TABLES

Table Page

1 Reliability of Administrative Rating of Instructors 16

2 Correlation of Administrative Rating with Other Measures of Instructor Effectiveness 19

3 Correlations Between Ratings of Teacher Characteristics 22

4 Reliability of Peer Rating of Instructors 25

5 Correlation of Peer Rating with Other Measures of Instructor Effectiveness 26

6 Reliability of Student Rating of Instructors 29

7 Correlation of Student Rating with Other Measures of Instructor Effectiveness 32

8 Intercorrelations by Trait of Student Rating of Instructors 33

9 Correlation of Grades Received by Students with Their Rating of Their Instructors 35

10 Relationship of Teacher Factors to Student Rating 36

11 Relationship of Student Factors to Student Rating 38

12 Relationship of Self-Rating to Other Measures of Instructor Effectiveness 41

List of Tables (Cont.)

Table Page

13 Reliability of Various Methods of Observing Teaching Effectiveness 44

14 Reliability of Measures of Student Gain 56

15 Correlation of Measures of Student Gain with Other Measures of Teacher Effectiveness 58

16 Correlations Between A.C.E. Psychological Examination and Various Measures of Teacher Effectiveness 62

17 Correlations Between Various Psychological Examinations and Measures of Teacher Effectiveness for Groups of Teachers of 90 or More 64

18 Relation of Education to Instructor Effectiveness 67

19 Educational Qualifications of "Best," White, High School Teachers 69

20 Relation of Practice Teaching Grades or Ratings to Scholarship 72

21 Relation of Practice Teaching Grades or Ratings to Teaching Effectiveness in the Field 74

22 Relation of Scholarship to Teaching Effectiveness in the Field 76

23 Age and Experience as Related to Teaching Effectiveness 80

24 Teaching Experience of Military and Civilian Instructors in Air Force Technical Schools 83

25 Relation of Scores on Subject-Matter Tests to Measures of Instructor Effectiveness 85

26 Relation of Scores on Professional Information Tests to Measures of Instructor Effectiveness 86

27 Relation of Extracurricular Activities to Instructor Effectiveness 89

28 Relation of Scores on the Cooperative General Culture Test to Measures of Instructor Effectiveness 91

List of Tables (Cont.)

Table Page

29 Sex of Instructor as Related to Instructor Effectiveness 94

30 Relation of Scores on Measures of Teaching Aptitude to Teaching Effectiveness 97

31 Relation of Interest Test Scores to Teaching Effectiveness . . . 100

32 Relation of Ratings of Voice and Teaching Ability 102

33 Opinion Studies of Traits, Qualities, and Characteristics of Successful Teachers 109

34 The Five Most and the Five Least Important of 46 Teacher Traits as Ranked by Four Groups of Judges 110

35 Relation of Personality Measures to Measures of Instructor Effectiveness 115

36 Relation of Social Adjustment Measures to Measures of Instructor Effectiveness 118

IDENTIFYING THE EFFECTIVE INSTRUCTOR:
A REVIEW OF THE QUANTITATIVE STUDIES
1900-1952

INTRODUCTION

The equipments of modern warfare are highly technical. Successful prosecution of a war demands that thousands of young men be able to maintain and operate electronic and mechanical devices that are often extremely complex. Since these men, upon induction, do not have the skills and knowledges necessary to such tasks, the armed forces are required to establish substantial training programs aimed at making satisfactory technicians out of raw recruits.

Fast and effective training requires at its core skilled instruction. The problem of how to select personnel who can successfully accomplish this accelerated instructional job is thus crucial to the armed forces. Methods of training these potential instructors most rapidly and efficiently must also be developed. Research in the area of selection and training of instructors has, therefore, very high probability of payoff in terms of a more efficient military organization. The first step would appear to be that of determining what is now known concerning the problems involved.

While the research literature was being surveyed as background material it became apparent that a summary of the findings of the quantitative studies had potential value for anyone concerned with instructor selection and training problems, not only in the Air Force, but also in the other services and in civilian institutions, schools, and colleges. With these wider implications in mind, a comprehensive and critical review of pertinent research reports has been prepared.

Over the past fifty years a considerable literature has been built up concerning the problems associated with teacher effectiveness. Many of the articles that have appeared merely reflect expressions of opinion in the form of "armchair" analyses of teaching. Others, often written by the original investigators, deal with theoretical considerations arising out of research studies. Undoubtedly many of these general discussions are worthy of attention. Inasmuch as the more pregnant theoretical implications usually form an integral part of reports of actual research investigations, it was decided to include in this review only those studies that involved a quantitative attack on problems concerned with teaching effectiveness. Some exception was made in the case of a few of the most recent theoretical discussions by leading investigators in the field. Limiting the scope of the review in this manner reduces the bulk of material to be handled without seriously limiting the analyses of the problems of assessing teaching effectiveness or neglecting the progress that has been made in solving these problems.

In the search for quantitative studies over 900 references were examined. Of these, over 360 were abstracted for inclusion in the review. To obtain these references the following sources were used: Educational Index, Psychological Abstracts, and some 40 reviews and bibliographies, including the comprehensive Domas-Tiedeman (380) bibliography. A selected list of 28 of these reviews and bibliographies is included with the references accompanying this report. While no assurance can be given that all important research on instructor effectiveness has been covered, the reviewers had available the extensive facilities of the library of the University of Illinois as well as other sources of information.

Findings are presented as given in the original reports, even though in some cases the research designs are obviously faulty, or insufficient numbers of subjects have been used to allow statistically significant generalizations to be drawn. The discussions of research studies and the tabular material are presented chronologically under each topic heading, except in a few instances where some specific feature of the investigations is emphasized (e.g., in Table 30 order of presentation is chronological for each test). The chronological order enables the reader to judge results in terms of the tendency in later work to use more precise statistical methods, improved research designs, and to report more meticulously the conditions under which an experiment was conducted.

An attempt has been made to include in the tables all information considered necessary for interpretation of results. In the column describing the samples used in the various studies, besides the size of the sample, level of teaching position is stated wherever known. Other data on which a sample was selected are also given, such as: the sample was a dichotomous one of good-poor teachers, or, it was composed of only inexperienced teachers. In cases where this additional information is not included, it may be assumed that the sample was indeterminate except for the particular variable cited.

From the arrays of results that have been assembled, the reviewers have set down what appeared in their opinion to be the most probable generalizations arising from the data and have drawn certain conclusions from these to serve as a guide in Air Force technical training research projects. It is anticipated that these facts and conclusions may also assist other investigators in research planning in this field.

PRINCIPAL FINDINGS OF CITED RESEARCH STUDIES

Criteria

The main findings of the quantitative studies reviewed in the present report will be summarized.

Surveys of rating devices. Surveys of appointment blanks and rating scales in use have failed to provide means for identifying the significant items to be used in setting up instructor rating devices. The most frequently mentioned qualities on existing teacher appointment blanks are ability to discipline, ability to teach, scholarship, and personality. There is no general agreement as to what constitutes the essential characteristics of a competent teacher. Similarly, items on present rating scales tend to be subjective, undefined, and varied, there being no consistency as to what traits a supervisor might be expected to observe and evaluate.

Administrative ratings. Administrative over-all opinion constitutes the most widely used measure of instructional competence. Available studies show in general that teachers can be reliably rated by administrative and supervisory personnel (usually with *r*'s of .70 or above). For the most part, administrative ratings do not produce very high correlations with measures of student gain. Intercorrelations of rated traits or categories appear to give evidence that traits which are more objectively observable or are more independent of opinion tend to be less prone to logical error or halo effect than are those traits which are more intangible and hence more subjectively estimated. The implication seems clear that by and large ratings made by the same person are apt to be contaminated by halo and that in many such instances a single rating of over-all effectiveness may be as useful as an evaluation based on a composite of a number of ratings of separate traits.

Peer ratings. Peer ratings have been little used. For administrative purposes they are probably not too useful since teachers have certain misgivings about passing judgment on fellow teachers. From a research standpoint in using peer opinion, ranks will probably give better results than ratings. There is considerable agreement between supervisors and fellow instructors in ratings of instructors. As in the case of administrative ratings, considerable correlation is found among ratings given different traits by the same peer raters. That is, halo influences peer ratings just as it does administrative ratings.

Student ratings. The use of student ratings of instructor effectiveness appears to be growing. Such ratings tend to show fair consistency, their reliability, as with other ratings, increasing with the number of ratings pooled in fairly good accordance with the Spearman-Brown formula. When student ratings have been compared with other measures of instructor effectiveness, rather diverse results have been found depending in part upon the criteria employed. Considerable halo effect is usually found when students rate their instructors on several traits. Whether or not grades received by students affect their ratings apparently depends upon the instructional situation. Results may indicate that if the instructor favors the brighter students he will be approved by them and a positive correlation between student ratings and grades will result. If he teaches for the weaker students he will be disapproved by the brighter students and a negative coefficient will be obtained. By and large such factors as size of class, sex of students, age or maturity of students, and intelligence or mental age of students seem to have little bearing on student ratings. Research has been too

POOR ORIGINAL COPY - BEST
AVAILABLE AT TIME FILMED

sporadic and results too inconclusive to allow generalizations to be made concerning the influence on student ratings of other factors such as age and sex of teacher, length of students' acquaintance with the teacher, length of time teacher has taught in the school or taught a student, pleasurable personal relationships between student and teacher, and whether or not subject taught by rated teacher is students' favorite subject. There is considerable expressed opinion but little research evidence that student ratings will contribute to instructor improvement or could be used to improve supervisory ratings.

Self-ratings. While there is some tendency for instructors to overrate themselves, self-ratings show negligible relationship with administrative ratings, student ratings, or measures of student gains. On the basis of the few available studies of self-ratings of instructors, the obvious, undisguised self-rating technique would seem to offer little encouragement for evaluative or research purposes.

Systematic observations. Systematic observation techniques to determine differences in performance of effective and ineffective instructors have been largely neglected in research in the instructor area. Most of the observations made have been dependent upon the subjective judgment of the observer. In general, the reliability of planned observational recording compares favorably with other methods of instructor evaluation. The most general criterion of validity of observation has been face validity. No single, specific, observable teacher act has yet been found whose frequency or per cent of occurrence is invariably significantly correlated with student achievement. There seems to be some suggestion, however, that questions based on student interest and experience rather than assigned subject matter, the extent to which the instructor challenges students to support ideas, and the amount of spontaneous student discussion may be related to student gains. Apparently there are no optimum time expenditures for particular class activities; a good instructor may function successfully within a wide range of time expenditures. A factor analysis of a number of instructor and student behaviors resulted in three factors: (a) understanding, friendliness, and responsiveness on the part of the instructor, (b) systematic and responsible instructor behavior, and (c) the instructors' stimulating and original behavior.

Student gains. Of the several methods used to measure student change, residual student gain, that is, the difference between actual gain and predicted gain, is becoming more widely used as a criterion of instructor effectiveness. With all its difficulties it appears to offer one of the best criteria thus far used. As compared with commonly reported test reliability coefficients those obtained in gains studies have been low. The great discrepancies in the findings of investigators who have examined the student gains criterion emphasize the extreme variability in relationship with other criteria used to indicate instructor ability. Within the limits of measures so far used, the relationship between administrative opinion of an instructor's competence and the amount of subject matter that the instructors will impart to his students cannot be predicted.

Predictors

Intelligence. Whether or not intelligence is an important variable in the success of the instructor apparently depends upon the situation. In general there appears to be only a slight relationship between intelligence and rated success of an instructor. Correlation coefficients for high school teachers tend to be somewhat higher and somewhat less variable than those reported for elementary teachers. For all practical purposes, however, this variable appears to be of little value as a single predictor of rated instructor competence.

Education. Considered as a group, the investigations of semester hours or years of education as related to instructor efficiency have indicated that any relationship that may exist is slight. Beyond certain more or less obvious knowledge requirements, greater or lesser education of a teacher in terms of courses or semester hours seems to be unimportant in discriminating between good and poor teachers.

Scholarship. Implications of studies reviewed with respect to scholarship are quite clear. Grades a student will obtain in a practice teaching course may to some extent be predicted by the grades that student obtained in college. Accurate prediction of success in practice teaching, however, cannot be made on the basis of an individual's scholastic record in high school. Almost all available studies report low positive correlation coefficients between measures of on-the-job performance of teachers and earlier scholarship as reflected in over-all achievement in high school or college, or in standing obtained in specific college courses (including practical teaching courses). There appears to be some relationship, but it is small. No investigator has shown that the attainment of a particular standing in high school or college or the mastery of any single course or group of courses is essential to teaching competence. The positive correlation coefficients usually found probably reflect primarily the relationship of general intelligence to both academic and teaching success.

Age and experience. It appears that a teacher's rated effectiveness increases at first rather rapidly with experience and then more slowly up to five years or beyond. There is then a levelling off, and the teacher may show little change in rated performance for the next fifteen or twenty years, after which, as in most occupations, there tends to be a decline.

Knowledge of subject matter. Whether or not knowledge of subject matter is related to instructor competence seems to be a function of the particular teaching situation. Some studies suggest that too much knowledge on the part of the teacher may result in teaching "over the heads" of students.

Professional information. Scores on tests of professional information appear to bear some slight relationship to supervisory ratings or rankings of instructor competence. Contradictory results have been obtained, however, when such scores are correlated with pupil gain.

Extracurricular activities. In general, investigators have found low positive relationship between an individual's participation as a student in extracurricular activities and his later instructor effectiveness.

General culture. Studies reviewed appear to indicate that the relation of Cooperative General Culture Test scores to instructor effectiveness differs little from those reported for other subject matter tests.

Socioeconomic status. Studies of the relationship of socioeconomic status (as measured by such devices as the Sims Socio-Economic Scales) to criteria of instructor effectiveness show little, unless it is that those from higher status groups have greater probabilities of success in life than those less fortunate.

Sex. No particular differences have been shown when the relative effectiveness of men and women teachers has been compared.

Marital status. Despite some prejudice to the contrary there appears to be no evidence that married teachers are in any way inferior to unmarried teachers.

Teaching aptitude. Results obtained from measures designed to predict teaching ability show great disparity. Data thus far available either fail to establish the existence of any specific aptitude for teaching with any degree of certainty or indicate that tests used were inappropriate to its measurement.

Teaching attitude. Attitude toward teachers and teaching as indicated by the Yeager Scale devised for its measurement seems to bear a small but positive relationship to teacher success measured in terms of pupil gains.

Interest in teaching. In most of the studies reviewed, interest in teaching was measured by interest test scores which indicated similarity of interest of teachers and persons undergoing the interest test. Correlations resulting from the use of several standard interest tests either cluster around zero or are so inconsistent as to render such tests of rather doubtful value as predictors of teaching success. The common factors that were found through factor analyses to underlie the reasons given for choosing the teaching profession are perhaps provocative of further research but were based on too few cases to justify any clear-cut interpretation.

Voice and speech characteristics. On the basis of studies reviewed, in general, it appears that the quality of the teacher's voice is not considered too important by school administrators, teachers, or students. In one study, however, certain speech factors were found to be correlated significantly with student gains and with effectiveness ratings of supervisors. The intercorrelations of the speech factors, however, were so high that general speech ability based on a single factor is probably as useful as a composite of judgments based on several speech factors.

The photograph. Studies of the use of the photograph as a predictor of instructor effectiveness have failed to demonstrate that photographs have any predictive value.

Statistical analyses of instructor abilities. Such instructor factors as empathy, professional maturity, general knowledge, mental ability, social adjustment, and the like have been identified through factor analyses by various investigators. The statistical analyses so far reported, however, suffer from inadequacies of criteria, testing instruments, or number of cases.

Opinion studies of instructor personality characteristics. The attempts made to identify characteristics of successful and unsuccessful instructors by making lists of traits based on opinion appear largely sterile in terms of usability for evaluation or selective purposes.

Causes of teacher failure. In most of the studies of unsuccessful teachers poor maintenance of discipline and lack of cooperation tend to be found as the chief causes of failure. Health, educational background, training, age, and knowledge of subject matter, on the other hand, appear to be relatively unimportant factors in terms of teacher failure.

Personality tests. Results obtained with personality tests of teachers have shown wide variation when correlated with other measures. Some so-called personality tests appear to show significant correlations with certain measures of instructor effectiveness. Until carefully controlled, well-designed studies employing adequate numbers of instructors have been made, however, the problem of determining the personality patterns of effective teachers must still remain unsolved.

CRITERIA OF INSTRUCTOR EFFECTIVENESS

By common definition a criterion is any standard used for judging. For the scientist, however, such a definition is inadequate. A criterion which is to be used for scientific judgments cannot be just any standard. It should be the best possible standard for the particular class of judgments that are to be made. This means that the scientist must be able to justify his choice of a criterion by demonstrating its logical relevance to the problem at hand and by showing that it possesses measurement characteristics which are technically adequate.

So long as the investigator restricts his research to laboratory studies the establishment of a justifiable criterion usually presents no great difficulties. A criterion for memory, for instance, may be the recitation without error of a list of nonsense syllables, or the criterion of learning may be a specified minimum of blind alleys a rat enters while traversing a maze. The moment research is moved into less rigidly controlled life situations, however, the investigator is confronted with criterion problems which are seldom simple and often impossible of completely adequate solution. The determination of a scientifically justifiable criterion of instructor effectiveness presents such problems.

Every educational system and every training program has certain goals. The first requirement for choosing a criterion of instructor effectiveness is that these goals be defined. The measure of a particular teacher's effectiveness is then the extent to which that teacher facilitates the students' progress toward these goals. Since in any system there are usually several educational goals, a measure appropriate to each goal is indicated. The construction of a single, over-all criterion of instructor effectiveness would require that these various measures should be weighted into this criterion in accordance with supportable value judgments as to their relative importance.

Obviously, the fulfilling of the requirements for such a criterion of instructor effectiveness is a large order. The comparative student changes that would require measurement in certain educational systems, or at certain stages in a particular curriculum, might quite defensibly include such aspects as: changes in knowledges of specific subject matter, improved success in subsequent schooling, improved personal adjustment, or increased success in life. It is conceivable, also, that the effective teacher contributes to changes in other teachers' pupils through individual guidance, assistance in planning the school program, good influence on group morale, and the like, thus creating effects that cannot be isolated or ascribed to any one teacher.

In the studies reviewed, the criterion problems have been handled with widely varying degrees of sophistication. Measures found acceptable as criteria of instructor effectiveness by one investigator are often considered as unvalidated potential predictors by others. In order to provide for comparisons among studies and for appraisal of research progress, the reviewers have grouped together what appeared to them to be comparable studies. The basis for these groupings rests on the use by the investigators of similar criteria, or where no measures appeared to merit designation as a criterion, of similar potential predictors.

The largest grouping covers studies in which ratings or rankings of teachers have been used as criteria. Most commonly the reporting investigator does not deal explicitly with the problem of the relevance of such criteria to teacher effectiveness. In the opinion of the reviewers, if one is concerned with teacher effectiveness as the changes brought about by the teacher in the teacher's own pupils, then ratings and rankings are less relevant than either measures of student change or controlled observations of student behavior. Ratings are someone's estimate of the effects on students of those teacher characteristics the rater happened to observe, and which he deemed important. Without demonstration that these estimates have relationship to student achievement, they cannot really be considered as satisfactory substitutes for measures of pupil change. On the other hand, if one is considering that part of teacher effectiveness which the teacher contributes to the growth of all pupils by participation in the efforts of the educational group, then ratings or rankings would seem to be somewhat more relevant. In this latter case the influence of the teacher is a function of the quality of the teacher's relations with students in general, with other teachers, supervisors, and the community. Differential effectiveness is a matter of differential contribution to the over-all goals of the school or educational system. Since such contribution is almost inevitably in a cooperative setting, and since its effects are diffuse and (almost certainly) unmeasurable, there would appear to be logical justification for an attempt to get estimates of effectiveness in this area by the use of ratings or rankings obtained from other people in the educational situation.

Another section covers studies in which observational measures of teacher performance have been used. It is plausible that changes in students should be related to what the teacher does and how he does it. Furthermore, it seems reasonable that careful and objective observation of the teacher's behavior in the teaching situation could provide a measure of the teacher's effectiveness. A number of investigators have thus attempted to achieve objectivity in a criterion by the use of observational measures of teacher performance. However, before any method of objectively evaluating effective performance on the part of a given teacher can become useful, such method must be proved to be capable of measuring kinds of teacher behavior related to the type and amount of change the teacher produces in her pupils.

Studies that used measures of pupil change as a criterion are also grouped together. Granting that many of the pupil changes that would indicate a teacher's effectiveness are in behaviors that are not measurable, or at least have not yet been measured, there is at least one area in which measurements have been made. This is the area of student changes in knowledge of subject matter. While adherents of various educational philosophies might disagree as to the importance of changes in subject matter knowledge relative to other kinds of desired changes, it seems probable that all would agree that such changes have some importance and that they are relevant to the problem of teacher effectiveness.

The last section of the review covers other instructor or student variables or measures that were included in the studies read. These the reviewers have classified as "possible or potential predictors" regardless of what they were designated by the original authors. They are so classified because many of them, if their correlation with an adequate criterion of instructor effectiveness could be demonstrated, would be useful in the selection of personnel for teacher training or for assignment to teaching positions. Within the section the various classes of potential predictors or correlates are placed together to allow comparisons to be made and, where possible, conclusions to be drawn.

Rating the Effectiveness of Instructors

An Appraisal of Instructor Rating Methods

In attempts to evaluate instructors systematically many kinds of rating methods have been used (361). In a great many of the studies reviewed, investigators adopting rating as a criterion of teaching effectiveness have accepted the rating scale or method in use in a particular school situation. The types of rating scales which have been most favorably received by school administrators are the graphic, the check list, and to a lesser extent the rank order or order of merit. Consequently, these scales account for nearly all of the studies using rating as a criterion. In a few studies, however, the paired-comparison, critical-incidents, or forced-choice type of rating scales have been used.

The reason for the varying degrees of popularity of the different types of rating scales for administrative use is obvious. Ease of administration plus assurance that the administrator can follow his subjective leanings appear to have been the factors given the greatest weight in the choice of a rating method.

Since the results obtained in rating teaching effectiveness depend in part on the adequacy of the methods used, a brief appraisal of some of the more usual methods that have been applied to instructor rating seems appropriate to the purposes of this review.

The graphic rating scale is simple, comprehensible, easy to administer, free from direct quantitative terms, and discriminates as finely as the rater desires. It is also very susceptible to leniency effects.

The check list, on superficial appraisal, appears to be a simply constructed device though it is cumbersome to administer. To achieve a technically sound instrument, however, it is necessary to do more than just compile a collection of random statements. A thorough job analysis should be undertaken and as with other rating methods, comparative evaluation must be made of the various behaviors to discover those elements which determine good and poor instructors.

The rank-order technique while offering a simple means of evaluating instructors, lacks the popular appeal of the above two methods. From a research point of view its chief drawbacks are that it does not indicate the magnitude of the differences between persons rated nor does it indicate the differences between groups. This device has sometimes been used to validate other methods.

Although some investigators have claimed that the paired-comparison technique tends to be more accurate than rank-order or rating-scale methods, it has lacked favor among administrators, first, because it is extremely time consuming and laborious especially when used in rating large groups and, second, because there is usually a very high correlation between paired comparisons and rankings. It is also somewhat more resistant to manipulation by the rater than rank-order or graphic rating scales. Some investigators have recommended this device as a criterion of validity against which less rigorous methods of rating may be checked.

The search for more stable rating methods has led to the development of the critical-incidents and forced-choice techniques. Of these the forced-choice technique appears to be the more promising. The unique feature of this technique is that it limits the rater's control of the final result of his rating, thus effectively reducing biasability (272). Limiting the rater's control helps also to counteract another weakness usually associated with rating, that is, the raters tendency to become more and more lenient with repeated ratings. Nonbiasability effectively minimizes the effects of this changing frame of reference on the part of the rater.

The critical-incidents method, devised by Flanagan (113, 114), was developed as a means of identifying the important and valid behaviors on which rating should be made. So far it has not shown much promise in the rating of instructors. Domas (104) and Jensen (167) in attempts to use this method in school situations have demonstrated, perhaps unintentionally, the principal weakness of the method. When the "critical incidents" have been collected some attempt must be made to organize them so that they may be used conveniently. The resulting categories appear, however, as a list of vague generalities which might have been jotted down without going through all the elaborate process of accumulating the incidents. After Domas had collected 1000 and Jensen had assembled 500 critical incidents, they found they were unable to fit them into categories except as they represented effective or ineffective behavior and so presented them in their reports. Charters and Waples (74), incidentally, encountered the same difficulty when they attempted to organize lists of characteristics essential for successful teaching. Another principal weakness of the critical-incidents technique is that it depends entirely on the conception of effectiveness held by those who report the incidents. In applying the technique to teaching, its validity depends on the opinions of effective teaching held by the particular superintendents, teachers, students, or others from whose reports incidents are sought.

High reliability in terms of agreement among raters depends upon precise definitions of traits being rated so that raters have a common understanding of what is being rated, and sufficient frequency of occurrence of the behavior, trait, or quality so that systematic, extensive observations may be made. Wrightstone (361) reports studies by several investigators which tend to show that the following traits can be more reliably rated: efficiency, originality, perseverance, quickness, judgment, energy, scholarship, leadership, and intelligence. Such traits as courage, selfishness, cheerfulness, kindness, judicial sense, and tact proved not to be so reliably rated. These findings are perhaps specific to the raters, the rating scale used, the ratees, and the situation. It should be pointed out that it is doubtful if such literary traits as those exemplified here can be sufficiently well-defined to be useful nor can they be agreed upon by different raters, except perhaps as they uniformly reflect halo from an agreed on reputation. Asch (11) has shown that the content and functional value of a trait changes with the context of other traits. Gaining an impression of another person is not a process of fixing each trait in isolation and noting its meaning but rather a summation of the effects of these traits. For this reason it is probably more accurate to judge whole impressions rather than artificially isolated traits. Carefully planned studies, however, might well enable predictions to be made as to what types of traits and behaviors can be more reliably rated than others.

The reliability and validity of ratings tend to be reduced by several sources of error. Among these should be included judgments based on insufficient evidence, lack of training of the rater, and poor rating devices. Subjective rating scales depend largely upon memory and therefore are subject to errors by forgetting.

Another source of error lies in the fact that some raters tend to overrate and some to underrate, while still others tend to rate everyone near the middle of the scale. Thus, ratings made by different raters may reflect differences in rating habits rather than differences among the people rated.

Perhaps the greatest sources of error are those of "halo effect," first noted by Wells (349), and "logical error." Halo effect is the tendency of the rater to rate one trait or quality high (or low) because another trait or quality has been rated high (or low) or because the rater knows that the individual rated excels (or is particularly weak) in some respect. Logical error arises from presuppositions in the minds of the raters and lack of definiteness of the trait being rated.

Ratings also tend to become more and more meaningless with repeated use. This is well illustrated by the results of repetition of the same scale in rating Army officers. In 1922, 25% of Army captains were rated as excellent; by 1940 the percentage had reached 90%; while in 1945, 95% of captains received an excellent rating (13). Increased leniency with repeated ratings is probably not directly a function of the type of rating scale but rather due to the operation of social and situational pressures. With repeated ratings there tends to be a changing frame of reference on the part of the raters. It should also be noted that leniency tendency is not as serious a drawback under research conditions as contrasted with operational conditions.

The reliability of rating scales is increased by pooling the rating of several judges. As shown in findings reported by Bryan (61), Remmers et al. (270), and others, reliability in ratings increases with the number of ratings pooled in fairly good accordance with the Spearman-Brown formula. Bendig (29) in a study in 1952 of inter-judge versus intra-judge reliability of the order-of-merit method found the relationship between these two types of reliabilities to be U-shaped. The groups of judges with the most highly reliable and most highly unreliable intra-judge reliability showed the most group agreement. Furfey in Wrightstone (361) showed that reliability was increased also by subdividing traits and having ratings made on the subtraits.

In the next sections the results of studies dealing with ratings made by administrators, fellow teachers, the teacher himself, and students are reviewed. In interpreting the results of these studies the many sources of error in rating methods must be constantly borne in mind. By and large investigators have tended to ignore the problems of correcting for the various sources of error and have worked with ratings as though they were already a perfected criterion.

Surveys of Types and Content of Scales

In an attempt to determine what characteristics of instructors are considered desirable or essential by authorities in the field of education, several studies have been made of appointment blanks or rating scales as used by teacher-training institutions, university departments of education, or state departments of public instruction. In most cases the procedure consisted of collecting the forms used, tabulating the items on the rating sheets, and determining the total frequency a given trait or quality was mentioned on the rating devices used by all the institutions surveyed.

In 1920 Osburn (250) attempted to determine the desirable personal characteristics of the teacher by studying appointment blanks used by 121 teacher-training institutions. The outstanding finding of this investigation was the lack of agreement as to what constitutes the essential personal characteristics of a competent teacher. The universities tended to be in somewhat closer agreement than the normal schools. Ability to discipline, ability to teach, scholarship, and personality were the most frequently mentioned qualities.

A critical analysis of rating sheets in use for rating student teachers in institutions of the North Central Association of Secondary Schools and Colleges was made by Smith (318) in 1936. Of the 128 institutions replying to a request for information, 103 made use of some form of rating sheet. Approximately 77% of these depended solely upon personal opinions of the raters. In 1941 Samuelson (291) reported a survey of rating scales in use in approximately 50 teachers' colleges and schools in 29 states. The investigator's chief finding was the variety of practices and methods of measurement employed.

Graphic scales, usually with five-point scale division, predominated, although descriptive scales, letter scales, and numerical scales were also used. In 1940 Schellhammer (294) also examined rating procedures in 109 teacher-training institutions. The forms, he found, varied from a single rating of 9 items to a comprehensive scoring of 72 items on a seven-point scale. Intelligence and health items appeared most frequently, with no other item appearing more than 11 times. This seemed to indicate that there is no general agreement as to which characteristics the supervisor might be expected to observe and evaluate. Peterson and Cook (255) in 1930, Dean (99) in 1939, and Woellner (358) in 1941 also surveyed rating procedures used in teacher-training institutions.

Barr and Emans (19), in 1930, in order to determine what qualities are prerequisite to success in teaching, studied 209 teacher rating scales collected from cities of more than 25,000 population, from state departments of public instruction and from university departments of education. They reported that the 6939 items found in the rating scales tended to be highly subjective and undefined. The scales also varied widely in content and organization, many being either quite superficial or apparently representing special points of view or systems of teaching.

In 1945 Reavis and Cooper (262) surveyed rating methods in use in 123 city school systems. They reported that the most notable characteristic of the rating devices employed was their lack of uniformity. The instruments varied in type, in number of items to be rated, in specific characteristics included, and in individual responsible for the rating. In one city teachers were rated "only by degrees held." A total of 1538 items were included in the scales used. Of these only 256 appeared on more than one device.

It would seem that the survey method might provide an obvious way of determining the significant items to be used in setting up instructor rating devices. The studies summarized, however, appear largely sterile. The meaningless sort of results obtained are probably due to the failure of the surveyors to develop a rationale which could be imposed on the materials surveyed. The reliability of the categorizing of descriptive terms for traits or characteristics would have to be tested. Single judgments, or even judgments based on a group of closely associated judges, would not suffice. Rather, agreement should be tested for fitting the categories into the rationale by a series of independent judges, much in the same way that the reliability of the categorization of behaviors by independent observers is studied in time-sampling studies. Such surveys of content are not apt to produce results worth the effort until, through empirical or other means, hypotheses concerning what teaching characteristics should be rated are first formulated and then these hypotheses are checked by reference to institutional practice.

Types of Raters

Rating devices not only differ in form and content but they are also designed to be used by different classes of raters. An instructor's competence, for instance, may be rated by his supervisor or by an outside expert, by his fellow instructors, by his students, by himself, or by some combination of these. Most instructor ratings heretofore have been made by administrative personnel, but in recent years student ratings of their instructors have been receiving more and more widespread use.

Administrative Rating of Instructor Effectiveness

As has been repeatedly shown by surveys, many school systems employ unstructured rating procedures, the most widely used measure of an instructor's competence being the over-all opinion of the principal, supervisor, superintendent, or school inspector. On the basis of judgment of such administrative personnel, instructors may be selected, hired, promoted, or fired. To the best of the reviewers' knowledge, a rating form for teachers was first used administratively in Milwaukee in 1896 (170). By 1900 school systems in a number of other cities were also using rating forms.

Demonstrated lack of agreement among administrators, however, and the undependable nature of subjective opinions in general have led to frequent attempts to put instructor rating on a sounder footing through the use of more analytic administrative rating devices. One of the earliest attempts to quantify instructor behavior was the tentative scheme for the measurement of teaching efficiency outlined by Elliott (106) in 1910. He based his method on the premise that the teacher was an "octo-personality"--executive, projecting, supervising, professional-technical, social, physical, moral, and dynamic.

Investigations of the reliability, validity, and halo effect of administrative ratings utilizing rating devices will be examined in this report.

Reliability of Administrative Rating of Instructor Effectiveness

Reliability can be measured (a) between raters, (b) for a single rater from one rating scale or item to another (which may reflect halo effect), and (c) between ratings by the same rater from one occasion to another. The available studies appear to show that teachers can be reliably rated by administrative and supervisory personnel, the preponderance of reliability coefficients reported being .70 or above. As shown in Table 1, there is considerable variation, coefficients of reliability for rated general effectiveness ranging from .17 to .98. When traits or qualities other than general ability are rated, the reliabilities tend to be somewhat lower than those found for general effectiveness [Barr (16), Boardman (39)]. Part of the range of reliability coefficients can be ascribed

Table 1

Reliability of Administrative Rating of Instructors

Investigator	Teacher sample	Type of rating	Raters	Measure of reliability	Correlation
Hamrin (1927)	88 student	Graphic[a] (54 items)	Supervisor	Between two different raters	.32
Boardman (1928)	88 high school	Ranking (sigma position)	Supervisor	Split half (of raters)	.92
Jacobs (1928)	100 elementary (50 good, 50 poor)	Graphic (7 items) & ranking	Principal	Same rater, graphic vs. ranking	.54
		Graphic & general merit[a]	Principal	Same rater, graphic vs. general merit	.70
		Ranking & general merit[a]	Principal	Same rater, ranking vs. general merit	.56
Tiegs (1928)	77 elementary, 1 sem. experience	Graphic (12 items)	Supervisor	Rerating, same raters	.79
	75 elementary, 1 sem. experience	Graphic (41 items)	Supervisor	Rerating, same raters	.93
Barr (1928)	1 elementary, 2 yr. experience	Graphic (12 items)	Observer[b]	Rerating, same observer	-.16 to .77
		General merit	Supervisor	Rerating, same observer	.18
Almy & Sorenson (1930)	(Not reported)	Graphic (20 items)	Supervisor	Rerating, same rater	.92
	(Not reported)	Graphic (20 items)	Supervisor	Between two different raters	.72
	110 student	Graphic (20 items)	Supervisor	Split half	.92
Taylor (1930)	64 elementary	Ranking	Principal	Reranking, 1 yr. later	.65
	90 elementary	Rating & ranking	Educational specialist	Same rater, rating vs. ranking	.88 & .90
Bossing (1931)	165 high school	Graphic[a] (9 items)	Supervisor	Rerating, 1 yr. later, same rater 93 cases; different raters - 72 cases	.05
		General merit[a]	Supervisor	Rerating	.83
Broom & Ault (1932)	42, 46, 48 1 yr. experience	For State Department Education & for college placement	Supervisor	Between two different ratings	.14, 43, 48
Coxe & Cornell (1933)	(Number unreported) elementary, 2 yr. experience	Almy-Sorenson & unpublished scale	Supervisor	Same rater, different scales	.67
	422 elementary, 2 yr. experience	Unpublished scale	Supervisor	Rerating, 1 yr. later	.55
	112 elementary, 2 yr. experience	Almy-Sorenson Scale	Observer	Same scale, 2 different observers	.51
	112 elementary, 2 yr. experience	Torgerson Scale	Observer	Same scale, 2 different observers	.91
Barr, *et al.* (1935)	66 elementary	Ranking & rating (composite 7 scales)	Supervisor	Same rater, ranking vs. rating	.63
Hardesty (1935)	127	Graphic scale	Superintendent	Rerating, same raters	.51
Odenwaller (1936)	380 elementary	Ranking	Supervisor	Average intercorrelation between 3 raters (principal, assistant principal, & supervisor)	.76
	100 elementary	Ranking	Supervisor	Average *r*, principal vs. assistant principal	.71
	50 elementary	Ranking	Supervisor	Average *r*, principal and/or assistant principal vs. supervisor	.65
Sandiford, *et al.* (1937)	28	Graphic scale[a]	Inspector	Contingency coefficient (4 inspectors)	.80
		Rated from written comments	Inspector	Split half (7 inspectors)	.87
		Rated from written comments	College staff	Split half (7 raters)	.90
Brookover (1940)	12 high school	Purdue scale (10 items)	Superintendent	Between two different raters	.25
U.S. Office of Indian Affairs (1940)	209 all grades	Graphic (22 items)	Supervisor	Between two different raters	.74
Jones (1946)	54 high school	Two graphic scales	Supervisor	Same rater, different scales	.98
Galt & Grier	1141 flying instructors	Flying proficiency & teaching proficiency	Supervisor	Flying proficiency vs. teaching proficiency	.70
	198 flying instructors	Graphic (16 items)	Supervisor	Between two different raters	.65
Stolurow, *et al.*	85 gunnery instructors	General merit	Supervisor	Split half (8 raters, *N* = 85)	.94[e]
		Graphic (13 items)	Supervisor	Split half (8 raters, *N* = 85)	.76 to .92
	90 gunnery instructors	General merit	Supervisor	Split half (8 raters, *N* = 90)	.94
		Graphic (13 items)	Supervisor	Split half (8 raters, *N* = 90)	.79 to .93
	155 gunnery instructors	General merit	Supervisor	Between two different raters (*N* = 155)	.90

---

[a] Official rating. (In the Hampton study the first rating was the official rating received by the teacher at the end of the first year of teaching; the follow-up ratings were confidential, and the time lapse between the two ratings ranged from 6 months to 3½ years for the first follow-up and 1½ years to 4½ years for the second follow-up.)

[b] In this study the teacher was observed on two occasions by 60 visiting superintendents who were unacquainted with the teacher.

[c] 19 of the 61 teachers were rated the third time by the same superintendent who rated them the first and second times.

[d] Exclusive of 1- and 2-room schools.

[e] All Stolurow's coefficients were corrected by Spearman-Brown formula.

Table 1 (Cont.)

Investigator	Teacher sample	Type of rating	Raters	Measure of reliability	Correlation
Barker (1948)	60 (20 superior, 20 average, 20 poor)	14 personality items	Supervisor	Mean 2 raters vs. mean 2 others (14 *r*'s, *N* = 60)	.51 to .89
Brandt (1949)	7 elementary, 1 room	Rating on 3 different scales vs. rating on 4th scale	Superintendent	Reratings (8 yr. later)	.77, .73, .81
			Supervisor	Reratings (8 yr. later)	.42, .46, .71
	12 elementary, rural[d]	Rating on 3 different scales vs. rating on 4th scale	Supervisor	Reratings (11 yr. later)	.27, .19, .11
	22 high school	Rate & rerate same scale	Supervisor	Reratings (4 yr. later)	.48
	12 elementary, 1 yr. experience	Rate & rerate same scale	Supervisor	Reratings (4 yr. later)	.72
Hampton (1951)	59 elementary	Graphic (12 items)	Superintendent	Rating[a] vs. follow-up, same rater	.35 to .57
	64 elementary	Graphic (12 items)	Superintendent	Rating[a] vs. follow-up, different rater	-.15 to .17
	61 elementary[c]	Graphic (12 items)	Superintendent	Rating[a] vs. 1st follow-up	.24 to .45
	61 elementary[c]	Graphic (12 items)	Superintendent	Rating[a] vs. 2nd follow-up	.22 to .39
	61 elementary[c]	Graphic (12 items)	Superintendent	1st follow-up vs. 2nd follow-up	.15 to .41
	61 elementary[c]	General merit & graphic	Superintendent	Same rater, general merit vs. graphic	.34 to .77
	61 elementary[c]	General merit & paired comparison (14 items)	Superintendent	Same rater, general merit vs. paired comparison	-.34 to .51
	61 elementary[c]	Graphic & paired comparison	Superintendent	Same rater, graphic vs. paired comparison	.00 to .56
Lamke (1951)	32 high school, 1 yr. experience	Rating as to acceptability & graphic scale	Principal	Same rater, different scales	.80
Ryans (1951)	192 elementary	Rating "observer blank" & over-all rating	Principal	Same rater, different scales	.83
	165 high school	Rating "observer blank" & over-all rating	Principal	Same rater, different scales	.83
Bach (1952)	73 high school, 1 sem. experience	Rating "acceptability" & graphic	Principal	Same rater, different scales	.53
		Graphic	Supervisor	Same scale, 2 different raters	.55
Ringness (1952)	16 men, 1 yr. experience	Rating "acceptability" & graphic	Superintendent	Same rater, different scales	.90
	18 women, 1 yr. experience	Rating "acceptability" & graphic	Superintendent	Same rater, different scales	.82

---

[a] Official rating. (In the Hampton study the first rating was the official rating received by the teacher at the end of the first year of teaching; the follow-up ratings were confidential, and the time lapse between the two ratings ranged from 6 months to 3½ years for the first follow-up and 1½ years to 4½ years for the second follow-up.)

[b] In this study the teacher was observed on two occasions by 60 visiting superintendents who were unacquainted with the teacher.

[c] 19 of the 61 teachers were rated the third time by the same superintendent who rated them the first and second times.

[d] Exclusive of 1- and 2- room schools.

[e] All Stolurow's coefficients were corrected by Spearman-Brown formula.

to method. Where correlations were obtained when the same raters used different methods or scales, the coefficients tended to be high, e.g., the *r* of .98 is of this type. Where correlations were obtained between two different raters using the same method, coefficients were considerably lower, e.g., the *r* of .32 in the Hamrin study (143). Hampton (142) in her study of administrative ratings made in 1951 found that "correlations between successive trait ratings of the same persons were different from zero at the one per cent level, trait by trait, when the raters were the same and nominally equal to zero when the raters changed."

The reviewers found some confusion among authors as to whether reliability or validity was involved in certain of the correlation coefficients computed. When raters are of equivalent prestige, status, or standing, the reviewers have assumed that consistency of ratings, i.e., reliability, is intended. Such studies are reported in this section. When raters are of obviously unequal prestige, of different classes, or the comparisons are with an entirely different order of criterion variable, the studies are included in the following section on validity.

In order to insure that raters were confronted with a common situation, Shiels (308), in 1915, asked 110 principals to rate the same ten case studies of teachers for instruction and discipline on a five-point scale. Higher reliability would be expected than would be the case in rating real teachers since the judges were basing their opinions on identical data. The ratings, however, showed considerable variation and range, there being no instance of 100% agreement. There was, moreover, less than 75% agreement in all but four cases in rating instruction, and in all but two cases in rating discipline.

In Barr's study (16) similar results were found. When 60 visiting superintendents observed, for two different periods of 30 min. each, the teaching of one teacher relatively unknown to them and then rated the teaching effectiveness of that teacher, great divergence of opinion was found. The superintendents spread their ratings on all traits over at least 9 points of a 10-point scale and for more than 50% of the items over all 10 points. One superintendent commented on the poorness of the teaching, while another remarked that he wished he could employ the teacher in his school. Correlations between first and second observation by the superintendents also proved in general to be low. Barr stated that an outstanding fact brought out by this study was that supervisors cannot agree when asked to analyze a teaching situation about which they have no advance information. He concludes further that "conventional supervision is highly subjective."

Correlation of Administrative Ratings with Other Measures of Instructor Effectiveness

A number of investigators over the past thirty years have made comparisons of various criteria of instructor effectiveness. Their studies have been summarized in Table 2. The correlation coefficients, where reported, range from -.61 to .82, the former being determined by Jones (172) when he compared principals' ratings of 13 teachers with gains made by their pupils in English and the latter by Nanninga (238) when he compared principals' with assistant principals' ratings of 15 high school teachers. In some instances rather substantial coefficients were obtained when ratings of various types of administrators were compared [e.g., Brandt (51), Bryan (61), Nanninga (238), Tiegs (333)]. In these and other cases where relatively high correlations were reported, opportunities for collaboration, prior discussion, or other sources of contamination of data were not completely ruled out. For the most part administrative ratings do not produce very high correlations with measures of student gains [e.g., Brandt (51), Taylor (331)].

In Knudsen's and Stephens' (180) analysis of 57 published devices for rating teaching, they discovered that often the validity of the device was implied in the assumption of the competence of its designers to select significant traits. Forty gave no statistical evidence of validity or

Table 2

Correlation of Administrative Rating with Other Measures of Instructor Effectiveness

Investigator	Teacher sample	Measures compared	Correlation
Morton (1919)	71 high school, city	Inspector rating (4 categories) vs. salary	.16 to .28
	151 high school, rural	Inspector rating (4 categories) vs. salary	.26 to .39
Hill (1921)	135 elementary, composite of 3 cities	Administrator rating vs. pupil gain	.45
	135 elementary, each of 3 cities	Administrator rating vs. pupil gain	.45, .24, .19
Knight (1922)	39 elementary & high school	Supervisor rating vs. student rank	.74 (av.)
		Supervisor rating vs. peer rating	.96
Boardman (1928)	88 high school	Supervisor vs. fellow teacher ranking	.68
		Supervisor vs. student ranking	.56
Nanninga (1928)	15 high school, 5 years or more in same school	Principal vs. 9 graduate student ratings	.47
		Assistant principal vs. 9 graduate student ratings	.76
		Average principal, assistant principal vs. 9 graduate student ratings	.56
		Principal vs. assistant principal rating	.82
Tiegs (1928)	25 elementary, 1 sem. experience	Supervisor vs. principal rating	.70
Baird & Bates (1929)	470 elementary	Supervisor rating vs. pupil gain	.14
Light (1930)	28 high school	Supervisor rating vs. student ranking	"Close agreement"
Taylor (1930)	89 elementary	Principal vs. research department ranking	.51
	101 elementary	Two principal ranking vs. 2 research department rating	.63
	77 elementary	Principal 1923-4 ranking vs. research department 1924-5 rating	.45
	105 elementary	Estimated teaching ability vs. pupil arithmetic gain	.02 to .15
	105 elementary	Estimated teaching ability vs. pupil reading gain	.05 to .24
Simmons (1932)	40 elementary	Rating by 3 supervisors vs. pupil gain	"No relationship"
Coxe & Cornell (1933)	112 elementary, 2 yr. experience	Supervisor vs. 2 observers (Almy-Sorenson)	.37
		2 observers (Almy-Sorenson) vs. supervisor (10 points)	.48
		2 observers (Torgerson) vs. supervisor (10 points)	.41
Greene (1933)	32 college	Ranking by dean vs. ranking by peers and students	All but one of dean's 1st 10 & two of peers' 1st 10 same as students' 1st 10; all but two of peers' 1st 10 on deans' list
Odenweller (1936)	560 elementary	Superintendent, principal, & assistant principal (teaching effectiveness vs. peer ranking on personality)	.53
Bimson (1937)	25 high school	Supervisor rating vs. pupil gain	"No relationship"
Bryan (1937)	22 sr. high school	Superintendent vs. assistant principal (general teaching ability)	.42
	41 jr. high school	Superintendent vs. assistant principal (general teaching ability)	.65
	22 sr. high school	Superintendent vs. assistant principal (11 items)	.26
	22 sr. high school	Superintendent vs. principal (general teaching ability)	.23
	22 sr. high school	Superintendent vs. principal (11 items)	.02
	22 sr. high school	Principal vs. assistant principal (general teaching ability)	.32

Table 2 (Cont.)

Investigator	Teacher sample	Measures compared	Correlation
	22 sr. high school	Principal vs. assistant principal (11 items)	.21
	41 jr. high school	Principal vs. assistant principal (11 items)	.52
	22 sr. high school	Average administrator (general merit) vs. student rating	.36
	41 jr. high school	Average administrator (general merit) vs. student rating	.53
	22 sr. high school	Average administrator (10 traits) vs. student rating	.36
	41 jr. high school	Average administrator (10 traits) vs. student rating	.55
Cooke (1937)	9 to 48 elementary & high school	Composite principal vs. teacher self-rating	.02 to .49
Brookover (1940)	33 high school	Supervisor vs. pupil rating	-.08
	12 high school	Average 2 supervisor vs. pupil rating	.63
Porter (1942)	27 student teachers	Supervisor rating vs. pupil rating	"Close agreement"
Brookover (1945)	66 male high school	Supervisor rating vs. pupil rating	"Slight positive"
LaDuke (1945)	31 elementary, 1-room	Composite supervisor rating vs. pupil gain	-.24 to .06
Gotham (1945)	57 elementary, 1- & 2-room	Composite administrator (3 scales) vs. pupil gain	.40
		Composite administrator (2 scales) vs. pupil gain	.40
Lins (1946)	58 high school women, 1 yr. experience	Composite supervisor rating vs. pupil rank	.28
		Composite supervisor rating vs. pupil gain	.19
Jones (1946)	14 high school	Official vs. supervisor rating	.59
	7 high school	Official rating vs. pupil gain	-.27
	13 high school (English)	Principal rating vs. pupil gain (English)	-.61
	(Number unreported) high school	Principal rating vs. pupil gain (all subjects)	-.28 and .10
Von Haden (1946)	58 high school women, 1 yr. experience	Supervisor vs. interview (8 qual.)	.16 to .51
		Supervisor vs. interview digest (8 qual.)	.12 to .44
	39 high school women, 1 yr. experience	Supervisor vs. autobiography (8 qual.)	.17 to .52
Galt & Grier (1947)	1141 flying instructors	Two supervisor rating vs. student rating	.08, .06
Riesch (1949)	22 elementary	Superintendent rating (5 factors) vs. pupil gain	.20 to .81[a]
		Superintendent rating (5 factors) vs. pupil gain	.13
Brandt (1949)	9 elementary, 1-room	Supervisor rating (8 yr. before) vs. pupil gain	.14
	17 rural	Supervisor rating (11 yr. before) vs. pupil gain	.35
	9 to 16 high school	Supervisor (M Blank) vs. three official rating scales	.45, .53, .65
	19 elementary, 1 yr. experience	Supervisor (M Blank) vs. industrial commission scale	-.02
Highland & Berkshire (1951)	635 AF technical schools	Supervisor vs. average peer rankings	.81
Lamke (1951)	32 high school, 1 yr. experience	Principal over-all vs. observer ratings	.69
		Principal vs. observer (same scale)	.61
Bach (1952)	73 high school 1 sem. experience	Principal "acceptability" vs. supervisor graphic	.39
		Principal graphic vs. 2 supervisor graphic	.48 & .44

---

[a] Pupil gain in attitude.

reliability; 11 mentioned correlations between ratings of the same teacher by different judges; 4 quoted correlations of weights assigned to various items on a given device by different judges; 3 gave correlations between successive judgments of the same judge; 2 included intercorrelations of scores assigned by different judges; and 2 mentioned correlations of scores on items with scores on general merit.

Intercorrelations of Rated Traits or Categories

Some, if not all, of the studies reported in this section appear to give evidence of the presence of the halo effect which tends to bias ratings in general. A number of the investigators whose studies are reviewed here have called attention to this factor as at least a partial cause of the large correlation coefficients found when ratings of several traits by the same rater are compared. Other investigators report high correlations without comment.

In any interpretation of these studies it is important to recognize that, for instance, a coefficient of .90 between rated "efficiency" and rated "use of methods" does not mean that good methods lead to efficiency; it merely means that raters tend to rate a given person at the same relative level on the two traits. It should be noted that two kinds of intercorrelation may indicate halo effect. The first kind is the correlation found when ratings of two traits by the same rater are plotted against each other. The second kind of correlation is that found when mean ratings of two traits of several instructors are plotted against each other.

In Table 3 12 studies are summarized in which correlations were computed [in the Bryan (61) and Brookover (55) studies actual coefficients were not reported] between some rating of general teaching merit and ratings on some other teaching characteristic where the two types of ratings were made by the same rater. It will be noted that, in general, the coefficients tend to be high, probably indicating operation of considerable halo effect. In some cases the relationships are quite as ridiculous as those Knight (178) found and commented on in his study of peer ratings. Knight obtained a correlation coefficient of .94 between general teaching ability and intellectual ability and one of .79 between teaching ability and skill in discipline when these were rated by fellow teachers. He also found a correlation of .86 between ratings on skill in discipline and intellectual ability. In pointing out the absurdity of these correlations, Knight said, "Were this really the truth, what a prodigy of intellect the 'strict,' but often dull, teacher would be!" Further, "If we thus generalized, we would also hold that Grant, admittedly a past master in control, also towered above Lincoln in mental stature."

In the case of certain traits, however, the correlation coefficients are low. For instance Ruediger and Strayer (283) report a coefficient of .04 between general merit and health and .20 between general merit and

Table 3

Correlations Between Ratings of Teacher Characteristics

Investigator	Teacher sample	Rater	Traits compared	Correlation
			General Merit vs.:	
Ruediger & Strayer (1910)	204 (approx.) elementary	Supervisor (18 to 26 school systems)	Keeping order	.56
			Teaching skill	.54
			Initiative	.50
			Health	.04
			7 other traits	.02 to .46
Boyce (1912)	343 high school	Supervisor (27 systems)	Instructional skill	.90
	343 high school		Rated pupil success	.85
	325 high school		Stimulation of pupils	.80
	311 high school		Health	.18
	376-330 high school		18 other traits	.36 to .71
Fordyce (1919)	123 student	Supervisor (12 systems)	Social power	.75
			Technique of teaching	.79
			Personality	.59
Baird & Bates (1929)	520 elementary	Principal	Control over method	.85
	571 elementary		Professional spirit	.84
	466 elementary		Social intelligence	.57
			4 other traits	.58 to .72
Odenweller (1936)	560 elementary	Supervisor	Personality	.83
Bryan (1937)	47 jr. & 22 sr. high school	Supervisor & pupils	10 other traits	Greater similarity in supervisor rating, item to item, than in pupil rating.
Brookover (1945)	66 male high school	Administrator	A number of traits	"Significantly correlated"
Galt & Grier (1947)	198 flying instructors	Supervisor	16 different traits	.26 to .90
Stolurow, *et al*. (1947)	132 gunnery instructors	Supervisor	Today's preparation	.83
			Ability to express self	.81
			Appearance	.35
			10 other traits	.54 to .79
Hampton (1951)	61 elementary	Supervisor (graphic scale)	Resourcefulness	.77
			Discipline	.70
			Health & vitality	.34
			9 other traits	.46 to .68
	61 elementary	Supervisor (paired comparison)	Resourcefulness	.51
			Knowledge of subject matter	.38
			Health	-.34
			9 other traits	-.33 to .35
			Pupil reaction vs.:	
Morton (1919)	71 high school, city	Educational specialist	Personality	.56
			Scholarship	.60
			Method	.81
	151 high school, rural	Educational specialist	Personality	.55
			Scholarship	.45
			Method	.66
Almy & Sorenson (1930)	110 student	Supervisor	Rating vs. grade assigned by same supervisor	.74

appearance; and Boyce (48) found a correlation of .18 between general merit and health. On the other hand Boyce reported a correlation of .90 between general merit and instructional skill. This suggests that traits which are more objectively observable or are more independent of opinion are less prone to logical error or halo effect than are those traits which are more intangible and hence more subjectively estimated. The implication seems clear that, by and large, ratings made by the same person are apt to be contaminated by halo and that in many such instances a single rating of overall effectiveness may be as useful as an evaluation based on a composite of a number of ratings on separate traits.

Peer Rating of Instructor Effectiveness

Apparently little use has been made of the practice of having teachers rate their fellow teachers. Roberts and Draper (279) in 1927 obtained material on the scope and character of the work of the principal from principals' reports from 441 high schools having an enrollment from 5 to 4000 pupils in all sections of the United States. Only 12 principals asked teachers to rate each other and 379 did not require such ratings; no answer to this question was given by the remainder of the principals. A survey made by Reavis and Cooper (262) in 1945 on rating methods in use in city school systems showed that in only two systems was teacher opinion used as part of the rating set-up.

In a number of studies, however, lists of desirable traits of teachers have been compiled by teachers themselves (53, 120, 173, 215, 303). A detailed analysis of these and other related studies is included in the section on Opinion Studies of the Personality Characteristics of Effective and Ineffective Instructors.

Superficially at least, the most obvious way to discover how a man does a job is to ask a fellow employee. It would seem that fellow-teacher opinion should provide a valid measure of instructor competence. The rating a teacher makes of a fellow teacher, however, is probably rarely based on first-hand observation but rests more often on hearsay and reputation. Even if he does have opportunity to observe other teachers' performance in the classroom, he may not know what is important to look for.

Furthermore, peer ratings have never been popular. This is probably due to the dislike of persons to evaluate or to be evaluated by their close associates. The raters can never be absolutely certain that uncomplimentary opinions do not get back to the person rated, nor are they always sure just how their ratings will be used. They are loath, for instance, to accept any responsibility for separating even an incompetent fellow worker from his job.

For administrative purposes, therefore, peer ratings of instructors are probably not too useful since teachers tend to have certain misgivings about passing judgment on fellow teachers. When obliged to rate their fellow teachers, they are apt to do what is popularly called a "snow job." They are careful to give only favorable ratings, thus avoiding any repercussions if their ratings became known to the one rated. This means of course that the instructor-rater keeps his more candid opinions to himself. From a research standpoint, in using peer opinion, ranks might give better results than ratings, especially if steps are taken to assure the raters of the anonymity of the results.

Reliability of Peer Rating of Instructor Effectiveness

Not many data were found on reliability of peer rating of teachers. Four studies in which reliabilities were obtained for fellow-teacher ratings are presented in Table 4. In these studies the N used was the number of instructors rated.

Correlation of Peer Rating with Other Measures of Instructor Effectiveness

Several investigators have been interested in showing the relationship between peer rating and other measures of instructor effectiveness. The rationale for making such comparisons appears to be that of lending support to the validity of the measure used in a particular study. Apparently there is considerable agreement in opinions of supervisors and fellow instructors. This would seem to indicate that the reputation of an individual is a common element in influencing the judgment of all who are associated with the teacher whether pupils, fellow teachers, or supervisors.

In the four available studies where correlations were computed, the coefficients ranged from .53 to .96. These four studies have been summarized in Table 5, together with three reports where noncorrelational methods were used in comparing peer ratings with other measures of instructor effectiveness.

Intercorrelations of Peer Rating of Instructor Effectiveness

As in the case of intercorrelations between traits rated by the same person for administrative ratings, close relationship is found for ratings given different traits by the same peer raters in the few studies available.

In 1922 Knight (178) in a study of 153 elementary and high school teachers found that mutual judgments of teachers with respect to general teaching ability correlated with their judgments of intellectual ability

Table 4

Reliability of Peer Rating of Instructors

Investigator	Teacher sample	No. of raters	Measure of reliability	Correlation[a]
Knight (1922)	153 elementary & high school	30	Split half (of raters)	.90[b]
Guthrie (1927)	16 college	5	Average of pairs of raters	.26
Boardman (1928)	88 high school	88	Split half (of rankers)	.72[b]
Odenweller (1936)	100 elementary	3	Average intercorrelation among the 3 raters[c]	.42
	49 elementary	3	Average intercorrelation among the 3 raters[c]	.41
	48 elementary	3	Average intercorrelation among the 3 raters[c]	.29

---

[a] The correlation coefficient is based on the teacher sample, the number of instructors rated.

[b] Uncorrelated.

[c] Ratings of personality.

Table 5

Correlation of Peer Rating with Other Measures of Instructor Effectiveness

Investigator	Teacher sample	Ratings compared with peer rating	Correlation
Knight (1922)	153 elementary & high school	Supervisor estimates (over-all)	.96 (mean r)
	39 elementary & high school	Pupils' estimates (ranking)	.58 (mean r)
Boardman (1928)	88 high school	Supervisor ranking (average sigma position)	.68
		Pupils' ranking (average sigma position)	.66
Greene (1933)	32 college	Ranking by dean	All but two of peers' 1st 10 on dean's list of 10 best.
Odenweller (1936)	560 elementary	Superintendent, principal, & assistant principal (teaching effectiveness) vs. peer ranking on personality.	.53
Albert (1941)	1 high school	1 administrator vs. 140 pupils and vs. 16 teachers (on a rating scale devised by author)	Pupils agree more closely with teachers than with administrator.
Ferguson & Hovde (1942)	1 high school	304 pupils & unknown N of peers rated 12 personality traits	Peers rated 5 higher, 4 same, 3 lower.
Highland & Berkshire (1951)	635 AF technical school	Average rankings of peers vs. rankings	.81

.94 and with judgment of skill in discipline .79, while judgments of skill in discipline correlated with intellectual ability .86. He concluded that in judging particular traits, "general estimate" (i.e., halo) influences the ratings to such a degree that judgments of particular traits are in themselves of little practical use.

Odenweller (247) also noted that in his study correlations are markedly higher when both traits are judged by the same set of judges than when one is judged by one set of judges and the other by another.

Student Rating of Instructor Effectiveness

In recent years certain educators have been quite voluble in advocating the use of student rating in evaluating the effectiveness of instructors. It is maintained that such ratings tend to raise standards of instruction by providing a basis for weeding out incompetent instructors and for improving the effectiveness of good instructors. These ratings, it is said, provide administrators with a means for securing dependable information which they should possess as to the opinions of students with respect to every member of the teaching staff.

That student ratings, within the limits of their reliability, are valid measures of student opinion of instructors cannot be questioned. It is probably true also that students being in a more or less close relationship with their instructors are in a better position than anyone else to make certain judgments of them. Whether or not these student ratings are in turn related to over-all effectiveness of the instructor in the teaching situation has not been demonstrated. There may be a closer relationship between pupils' success in school and their reaction to the teacher than there is between their success and methods of teaching or the so-called important physical aspects of the school environment and teaching aids.

While the practice of obtaining student ratings appears to be growing, their disadvantages have frequently been pointed out. Some administrators oppose them because of the cost in time or money or because of their possible disruptive effects upon student and staff morale. Among instructors there is considerable opposition to student ratings. Certain instructors fear the misuse of student opinion as a basis for advancement or separation of personnel. They point out also that student ratings may make instructors emotional, self-conscious, or resentful and that attempts to cater to student opinion may produce changes in undesirable directions. Students may lose respect for their instructors by being encouraged to set themselves up as judges of instructor competence. Instructors contend that student ratings are unreliable because of immaturity and prejudices of the raters who are influenced by grades, interest in specific subject matter, reputation of particular instructors, difficulty or ease of course material, and the like. Many students also are unfavorably disposed to rating their instructors. They consider such ratings a waste of their time unless administrative action results. Students themselves point out that the preferred instructor is often young, genial, and entertaining, while the serious, more experienced individual who stresses subject matter and insists upon certain standards of deportment and effort is rarely popular.

Quite a number of investigators have reported studies of student rating as a measure of instructor effectiveness and also as a means of instructor improvement. Among these are the studies of Bryan (61, 62, 63, 64, 65, 66), Starrak (322), Riley et al. (276), Goodhartz (131), and Remmers and his associates (264, 265, 266, 267, 268, 269, 321, 348). Galt and Grier (126) in a report of an investigation of flying instructors state that they found student rating useful and suggest that such ratings might well be looked into further. In a very recent study Flesher (115) has suggested that the question of whether or not ratings of an instructor might be inferred from their students' rating of the course taught by the instructor might well bear investigation. Flesher contends that student rating of courses tends to be more objective and frank and hence more valid than their ratings of instructors. In a limited test of this hypothesis done as a by-product of another study, Flesher obtained correlations ranging from .60 to .82 between course ratings and instructor ratings, with mean ratings for courses tending to be lower and more variable.

Reliability of Student Rating of Instructors

It might be expected that higher reliability coefficients would be obtained for composite student ratings than for composite administrator ratings of instructors because of the usually much larger numbers of student raters as compared with administrators making the ratings. As shown by the investigations summarized in Table 6, however, there is considerable variation in the reliability of student rating.

It will be noted that two kinds of correlational studies have been included in Table 6. In most of the studies the correlation coefficients are based on the number of instructors. This obviously is the proper N where reliability of students ratings in differentiating instructor effectiveness is required. In four studies, Remmers and Brandenburg (267), Root (281), Smeltzer and Harter (315), and Amatora (4), the reliability coefficients show the consistency with which the same students rate a particular instructor, using either the same or different rating devices. These studies give no information as to the reliability of student ratings with reference to the instructor differentiation problem since the N used is the number of student raters and not the instructors rated.

In addition to the studies reported in Table 6, a number of investigators have reported findings which have a bearing on the reliability of

Table 6

Reliability of Student Rating of Instructors

Investigator	Teacher sample	Student sample	Scale	Method	Reliability coefficient[a]
Knight (1922)	11 elementary	200 (approx.)	Pupil ranking	Chance half of student rating	.77
	15 high school	40 "most dependable" per teacher	Pupil ranking	Chance half	.52
	13 high school	40 "most dependable" per teacher	Pupil ranking	Chance half	.91
Guthrie (1927)	87 college	285 (approx.) (average 8.25 per teacher)	Average rank	Chance half of student rating	.79
	(Number unreported) college	365 freshmen	Rank	Chance half of student rating	.56
	101 college	85 advanced	Rank & graphic rating	Average rank vs. average rating	.61
	87 college	285 (approx.)	Rank	Reranking (1 mo. later)	.89
Remmers & Brandenburg (1927)	3 college	30; 33; 33	Purdue (10 items)	Reratings	.83, .50, .64[b]
Boardman (1928)	88 high school	4 schools	Rank	Chance half average ranks by student	.78
Root (1931)	1 college	200	Check list, 42 items	Rerating	.95[b]
Wilson (1932)	97 college	(Not reported)	Graphic rating (35 items)	Two student groups for each instructor (35 *r*'s?)	.65 to .88
Bowman (1934)	21 student	8 to 40 per teacher	Graphic rating (7 items)	Chance half student ratings	.91
	30 student	10 to 42 per teacher	Purdue (10 items)	Chance half each item (10 *r*'s)	.58 to .92
	30 student	10 to 42 per teacher	Purdue (10 items)	Average chance half of student ratings	.84
Remmers (1934)	57 student	(Not reported)	Purdue (presentation)	20 student rating (20 *r*'s)	.11 to .50
			Purdue (stimulation)	20 presentations student rating (20 *r*'s)	-.09 to .47
			Purdue (interest)	20 presentations student rating (20 *r*'s)	.02 to .35
	37 college	(Not reported)	Purdue (3 above traits)	20 presentations student rating (Average of 20 *r*'s)	.43, .35, .29
Smeltzer & Harter (1934)	5 college	12 classes	45 items (5-point)	Signed vs. anonymous ratings (12 *r*'s)	.63 to .79[b]
Bryan (1937)	22 sr. high school	600 (average 66 per teacher)	11 items (5-point)	Average chance half of ratings (11 *r*'s)	.69 to .91
	41 jr. high school	900 (average 71 per teacher)	11 items (5-point)	Average chance half of ratings (11 *r*'s)	.61 to .94
Heilman & Armentrout (1936)	46 college	17 to 121 per teacher	Purdue (10 items)	Fifth placement vs. rating by students	.75
	23 college	(Not reported)	Purdue (10 items)	Reratings (6 to 7 yr. intervals)	.69
Bryan (1941)	86 high school (16 schools)	30 per teacher	10 items (5-point)	Average chance half	.83 to .92

---

[a] Uncorrected reliability coefficients; correlation based on instructor *N* except at *b*.

[b] Correlation based on student rater *N*.

[c] Corrected to "equal that of 25 ratings."

Table 6 (Cont.)

Investigator	Teacher sample	Student sample	Scale	Method	Reliability coefficient[a]
Davenport (1944)	48 high school	1250 (approx.)	25 items (5-point)	Average of 48 pairs of ratings	.86
		1250 (approx.)	Pupil ranking	Average of 48 pairs of rankings	.95
				Student ranking vs. rating	.46
Cook & Leeds (1947)	100 elementary	20 per teacher	50 item (Yes-No-?)	Split half of average ratings	.93[c]
Galt & Grier (1947)	277 flying instructor	3.25 average each class	Graphic (18 items)	Class-to-class	.36
Amatora (1950)	(Not reported)	1174 elementary	7 intrascales	Graphic vs. Check List A	.74 to .88[b]
			7 intrascales	Graphic vs. Check List B	.75 to .88[b]
				Two check lists	.72 to .81[b]
			Total scores	Graphic vs. 2 check lists	.90 & .91[b]
			Total scores	Two check lists	.79
Drucker & Remmers (1951)	17 college	8 students (8 alumni, each institution)	Purdue (10 items)	8 student vs. 8 alumni (10 _r_'s)	.40 to .68

---

[a] Uncorrected reliability coefficients; correlation based on instructor _N_ except at _b_.

[b] Correlation based on student rater _N_.

[c] Corrected to "equal that of 25 ratings."

student rating but in which correlation coefficients are not reported. In 1926 Fritz (123) found that 89 students varied widely in their ability to duplicate their judgments on two ratings of one teacher obtained on a seven-part scale a week apart. In 1942 Porter (257) found, in having pupils rate some 27 student teachers, that some classes were considerably more lenient than others. Porter gave no statistical basis for his finding nor did he consider that the difference might be due to teacher merit rather than leniency of pupils if a teacher taught a better lesson in one class than in another. He concluded also that pupils tended to agree closely in judgments of best and poorest teachers but varied widely in their judgment of the middle group, a finding usually associated with the use of rating scales.

In 1929 Remmers (264), using the Purdue Rating Scale for Instructors, and in 1934 Starrak (322), analyzing ratings by students of the entire faculty of Iowa State University, reported that reliabilities obtained compared favorably with those of the best standardized objective tests. In 1932 Flinn (116) found that when an instructor was rated by four different supervisors and four different groups of pupils during a ten-year period the pupil ratings were much more uniform than were the ratings of supervisors. Flinn's result may simply reflect the fact that the standard error of an arithmetic mean is a function of the number of cases on which it is based and that a mean based on four different supervisors could fluctuate more widely than one based on a presumably larger group of pupils. In 1941 Albert (1) obtained consistent results when 78 high school teachers were rated by their 1578 pupils. In 1946 Remmers et al. (268) asked 559 engineers to use the Purdue Rating Scale in rating the best and worst instructors each had in college. The mean differences between best and worst instructors, as rated by the total group on the 10 traits of the scale and based on a total possible score of 100, ranged from 17.5 for personal appearance to 59.4 for stimulating intellectual curiosity. The average difference between means for the 10 characteristics was 39.6. These results are not too meaningful in the absence of standard deviations of the ratings of best and worst teachers.

Correlation of Student Rating with Other Measures of Instructor Effectiveness

A number of investigators have compared the results of student rating of instructors with those obtained from administrative and fellow-teacher ratings. Some have reported the obtained correlations as "validity coefficients." In a few instances, e.g., Lins (203) and Remmers et al. (269), pupil gain has been used as the criterion with which comparisons were made.

Table 7 summarizes 21 studies, in 12 of which correlation coefficients were reported. The considerable differences in magnitude of the coefficients obtained may be partly explained in terms of the diverse criteria employed, and in part they may be a function of the small numbers of teachers involved in most of the investigations. In general, the coefficients reported are quite high where ratings of teaching efficiency were used for both groups of judges. When a number of traits were rated, however, quite a wide range in coefficients resulted. This may have been due to the differing interpretation placed on the meaning of the traits by different raters. Results are not always comparable from study to study because of the lack of statistical controls. It was not always possible to tell from the reports, for example, when pupils ranked their teachers if corrections were made for size of groups. Knight (178) applied such correction, as did Boardman (39) who changed his ranks to sigma positions. Both got quite high correlations. Greene's study (135) which showed a high relationship between the teacher's salary and ranking by pupils may mean only that pupils were influenced by academic position.

Davenport (92) obtained a low correlation between teachers' self-ratings and pupils' ratings of teaching on comparable scales. He found a zero relationship between pupils' ranking of their teachers and the teacher's self-rating. Davenport suggests that a teacher's actual teaching may well be different from her philosophy of teaching, simply because such factors as size of class or other classroom factors force her to compromise.

It is interesting to note that in the two studies where pupil gain was one of the measures, only slight relationship was found. In the Lins' study (203) the low correlation might be due to the small number of teachers

POOR ORIGINAL COPY - BEST
AVAILABLE AT TIME FILMED

Table 7

Correlation of Student Rating with Other Measures of Instructor Effectiveness

Investigator	Teacher sample	Number students per teacher	Type of student rating	Other measures	Correlation
Knight (1922)	39 elementary & high school	40	Ranking	Peer rating (general merit)	.58
				Supervisor rating (general merit)	.74
Guthrie (1927)	(Number unreported) college	16.5	Ranking	Freshmen vs. advanced students	.52
Boardman (1928)	88 high school	(Not reported)	Ranking[a]	Ranking[a] by supervisor	.56
				Ranking[a] by peer	.66
Light (1930)	28 high school	44-287	Ranking	Supervisor rating	"Close agreement"
Greene (1933)	32 college	19	Ranking	Rank of salary	.86
				Peer ranking Ranking by dean	All but one of Dean's 1st 10 & two of peers' 1st 10 same as student's 1st 10
Bowman (1934)	30 student	8-40	Purdue Rating Scale	Critic teachers, same scale	-.50 to .47
Starrak (1934)	(Number unreported) entire college faculty	(Not reported)	Graphic (17 items)	Educational specialists	75% of students showed "close agreement"; 25% "divergence of opinion"
Bryan (1937)	22 sr. high school	20-152	Graphic (11 items)	Average 3 administrators (general merit)	.36
				Average 3 administrators (10 traits)	.36
	41 jr. high school	20-152	Graphic (11 items)	Average 2 administrators (general merit)	.53
				Average 2 administrators (10 traits)	.55
Brookover (1940)	33 high school	12-57	Purdue Rating Scale	Superintendent (same scale)	-.08
	12 high school	12-57	Purdue Rating Scale	Average 2 administrators	.63
Albert (1941)	1 high school	140	Graphic	Administrators (same scale) 16 peers (same scale)	Pupils agree more closely with teachers than with administrators.
Ward, *et al.* (1941)	40 student	(Not reported)	Purdue Rating Scale	3 supervisors (same scale)	-.09 to .996 median .88
Ferguson & Hovde (1942)	1 high school	304	Personality (12 traits)	Peer (same scale)	Peers rated five traits higher, four the same & three lower than pupils.
Porter (1942)	27 student	(Not reported)	Check list	Critic teacher rating	"Close agreement"
Davenport (1944)	44 high school	88.8	Graphic (25 items) "How Teachers Teach"	Self, "How I Teach"	.25
		80.1	Ranking	Self, "How I Teach"	.02
Von Haden (1946)	50 high school	5-6	Ranking	Educational specialist rating on basis of data collected in interviews & autobiographies	None of 27 *r*'s significant at .01 level. Range .06 to .34
Lins (1946)	58 high school	5-6	Ranking	Comparison 5 supervisors	.28
	17 high school			Pupil gain	.06
Galt & Grier (1947)	1141 flying instructors	3.12	Graphic (19 items)	Supervisors--Over-all flying proficiency	.06
				Supervisors--Over-all teaching proficiency	.08
Remmers, *et al.* (1949)	28 good; 25 poor (chem. lab)	(Not reported)	Purdue Rating Scale	Pupil gain	CR for 2 of 12 traits significant at .01 level; 4 at .02 level.
	28 good; 22 poor	(Not reported)	Purdue Rating Scale	Pupil gain	CR for no trait significant at .01 level; 1 at .02 level.
Cooper & Lewis (1951)	30 well-liked; less-liked by students	(Not reported)	Check list	"Absence of neurotic sign" Munroe Inspection Rorschach	.52[b] "No relationship"

---

[a] Sigma position.

[b] Tetrachoric correlation coefficient.

used in this part of the study or to some selective factor in the manner of choosing which students would rate each teacher. The traits on which differences were significant at the .01 level of Remmers' study (269) were: rating as compared to other instructors in the university and care of communal apparatus. Those significant at the .02 level were: supervision during tests and dailies, knowledge of chemistry, returning tests and dailies, should instructor be kept if suitable replacements are available.

Intercorrelations of Student Rating of Instructors

Ten studies in which intercorrelations were obtained between ratings by students for more than one trait are presented in Table 8. A divergence

Table 8

Intercorrelations by Trait of Student Rating of Instructors

Investigator	Teacher sample	Number students per teacher	Type of rating	Correlation
Remmers & Brandenburg (1927)	2 college	32	Purdue scale (10 traits)	-.02 to .62 .25 (average for all 10 traits)
Stalnaker & Remmers (1928)	1 college	94		-.07 to .72 .37 (average for all traits)
Remmers (1929)	115 college	(Not reported)	Purdue scale	.45 (average for all traits)
Boardman (1930)	87 high school	(Not reported)	Teaching efficiency vs.: Work hardest for Like best Discipline Learn most	 .73 .82 .75 .89
Bowman (1934)	21 student 30 student	8-40 (Not reported)	Seven traits Purdue scale (10 traits)	.12 to .79 .49 to .90
Remmers (1934)	64 student & 76 college	10	Presentation of subject matter vs. interest in subject Stimulating intellectual curiosity vs. interest in subject Presentation of subject matter vs. stimulating intellectual curiosity	-.005 & .18 .02 & .12 .12 & .19
Starrak (1934)	(Number unreported) entire college faculty	(Not reported)	Graphic (17 items)	-.06 to .63 .47 (average for all traits)
Heilman & Armentrout (1936)	46 college	17-121	Purdue scale Personal appearance vs. sympathetic attitude (lowest r) Stimulating intell. curiosity vs. presentation of subject matter (highest r)	.06 to .87 (31 of the 45 r's above .60)
Smalzried & Remmers (1943)	40 student	20-35	Purdue scale	.29 to .88 (28 of 45 r's above .60)
Henrikson (1949)	(900 ratings)	(150 total)	Effectiveness General merit vs. personality Voice merit vs. voice	 .66[a] .57[a]
Amatora (1950)	None--items on scale rated by students	(Not reported)	General rating vs. groups of items	.06 to .33 .51 to .66

---

[a] Coefficient of contingency.

POOR ORIGINAL COPY - BEST
AVAILABLE AT TIME FILMED

of results was evident in the various studies as to how much halo effect was present, even in cases where investigators used the same scale. Remmers and his associates (264, 266, 267, 321) in their several studies on the Purdue Rating Scale for Instructors show very little halo effect. As can be seen from Table 8 they reported consistently low correlations. In one study (321) only seven of the 45 intercorrelations proved to be above .50.

In the report of his study made in 1934, Remmers (266) says that his results emphasize the relative independence of the traits: interest in subject, Trait 1; presentation of subject matter, Trait 5; stimulating intellectual curiosity, Trait 10. In this study Remmers, in addition to the correlations reported in Table 8, determined halo effect by taking "five samplings of intercorrelations of five randomly selected pupils against five other pupils for Trait 1 versus Trait 5 and Trait 1 versus Trait 10." (Correlations were not computed between Traits 5 and 10 for some reason.) These were the 3 of the 10 traits appearing on the Purdue scale that were indicated by students as being the most important. Remmers averaged the r's without conversion to Fisher z's and without regard to the varying numbers of teachers involved in each r and then "corrected for attenuation." The resulting "true" correlation of .34, it seems to the reviewers, may be regarded with more than a little suspicion. In the case of college students, Remmers reported average r's corrected for attenuation of .52, .38, and .49 for Traits 1 vs. 5, 1 vs. 10, and 5 vs. 10, respectively.

In 1936 Heilman and Armentrout (148) also using the Purdue scale, found considerable halo effect and Smalzried and Remmers (314) in their factor analysis study of the Purdue scale, made in 1943, report that 28 of the 45 intercorrelations were above .60. Other investigators using different scales mention that quite a bit of halo effect was found. Bowman (47), in fact, in a third of a series of studies on student rating used an over-all rating because of the high intercorrelations among traits found in his first two studies.

Influence of Grades Received by Students on Their Rating of Instructors

The meaning of students' ratings of instructors is dependent to some extent on whether or not such ratings are related to grades received by students from the instructor concerned. If grades received are related to students' ratings, presumably instructors who gave high grades would be expected to receive higher ratings from their students than those who gave low grades. The presence or absence of the relationships here considered thus bears significantly on the validity assigned to students' ratings of their instructors.

The array of correlation coefficients presented in Table 9 is somewhat bewildering, particularly in the presence therein of coefficients

Table 9

Correlation of Grades Received by Students with Their Rating of Their Instructors

Investigator	Teacher sample	Number students per teacher	Type of rating	Academic measure	Correlation
Remmers (1930)	7 student & 4 college	16-32	Purdue scale (individual items)	Students divided into two groups on basis of grades	-.86 to .89 (biserial) -.71 to .45
Starrak (1934)	Entire faculty of one college	(Not reported)	Graphic scale (17 items)	Grades	.15
Bowman (1934)	9 student	8-40	12 characteristics	Grades	.004 to .65
				Difference between grade & student average grade	-.69 to .36
Smeltzer & Harter (1934)	5 college	(Not reported)	Graphic scale (45 items) Anonymous	Final examination	-.20 to .18
			Signed	Final examination	-.14 to .17
Krous (1935)	(Not reported)	(Not reported)	Analysis of "best" & "worst" teacher	Grade	No significant correlation
Heilman & Armentrout (1936)	46 college	17-121	Purdue	Teacher's severity of grading[a]	-.04
			"Fairness in grading"		-.24
Bryan (1937)	22 sr. high school 41 jr. high school	20-152	General teaching ability	Grades	.07 .15

---

[a] Obtained by computing the mean of all the grades assigned by each teacher for three quarters.

of substantial magnitude, but in both positive and negative directions. However, a hypothesis advanced by Remmers et al. (269) in 1949 makes such results plausible. These authors explain the apparently contradictory results obtained between this study and one by Remmers (265) at an earlier date in terms of methodology. In the earlier study the instructor was kept constant while students were varied in terms of grades and presumably scholastic ability. In the 1949 study the instructors were varied on the basis of whether or not their classes fell short or exceeded their predicted grade--presumably a measure of instructor ability. They point out that grades obtained under a single instructor and due to student differences may be either positively or negatively related to student ratings but that grades reflecting instructor differences rather than student differences are positively related to the ratings given instructors.

If one assumes that good students will approve of instructors who conduct their teaching at a high level (and over the heads of the poorer students), then, a positive correlation between student ratings and grades would result. Conversely, if the instructor pitches his teaching at the level of the weaker students, the brighter students will disapprove and a negative correlation will result. This hypothesis would account both for the range of coefficients obtained and for the fact that when correlations are not computed separately for each instructor, coefficients of negligible magnitude are found.

In those studies where grades were assigned "subjectively," i.e., where the instructor was directly responsible for the grade a student received, the relationship between grade and rating may reflect the students' response to the instructor's affective attitude. The relationship between student ratings and objective grades, on the other hand, may provide an indication of the students' reaction toward teaching competence. Another distinction among studies in this area is whether the correlation is between mean grades and ratings (where classes are the unit) as in the study of Heilman and Armentrout (148) or between individual ratings and grades (where the student is the unit) as in the report of Smeltzer and Harter (315).

Influence of Teacher Factors on Student Rating of Instructor Effectiveness

In addition to the grades a student receives a number of other factors have been investigated as having a possible influence on student rating of teachers. Among factors considered have been age and sex of teacher, length of students' acquaintance with teacher, length of time teacher had taught in the school or had taught pupil, pleasurable personal relationship between student and teacher, and whether or not subject taught by rated teacher was students' favorite subject. In view of the fact that research involving these factors has been rather sporadic and that some contradictory results have been reported generalizations cannot well be made. The few available studies are briefly summarized in Table 10.

Brookover (54, 55) in his two studies found what are apparently somewhat contradictory results. This might be explained by the fact that the measuring devices used by Brookover differed for the two studies. Brookover concluded that the nature of the pupils' personal relationships with their teachers affects their ratings of the teachers' abilities. This

Table 10

Relationship of Teachers Factors to Student Rating

Investigator	Teacher sample	Number students per teacher	student rating	Teacher factor	Relationship
Krous (1935)	(Not reported)	(Not reported)	Select best & poorest teacher	Taught student's favorite subject	Close relationship between favorite subject & subject taught by best teacher
Heilman & Armentrout (1936)	46 college	17-121	Purdue scale	Experience, age, & sex.	No reliable differences.
Brookover (1940)	37 high school	12-57	Purdue & person-to-person	Age & sex	No relationship
Davenport (1944)	51 high school	88.8	Graphic scale (25 items) "How Teachers Teach"	Number semesters student had been taught by teacher	No significant relationship
Brookover (1945)	66 high school male	(Not reported)	General merit	Age	Positive relationship
				Length of acquaintance with pupil	Positive relationship
				Length of time teacher had taught in school	Positive relationship
				Role in community	No relationship
			Pupil gain	Pleasurable personal relationship	Low, but significant negative

conclusion may be based on a form of halo effect, or more generally, a persistent response set on the part of the pupils. It is interesting to note that in the Brookover 1940 study, ratings of 39 teachers by their students on a scale measuring pleasant personal relationship yielded a correlation coefficient of .64 when correlated with superintendents' ratings. Boardman (40), in a study reported in 1930 in which pupils' rankings of teaching efficiency were correlated with their rankings of teachers in terms of for whom they worked hardest, the teacher liked most, the teacher having the best order or discipline, and the teacher from whom they learned most, found that when other factors were held constant pupils' liking for the teacher was the largest single factor in determining judgment of teacher efficiency.

In a longitudinal study of student ratings in which there was some turnover from year to year, Starrak (322), in 1934, found that rating scores of teachers tended to increase with successive ratings. This change was gradual, teachers originally placed in the lowest quarter moving to the second or third quarter by the end of a two-year period. Whether this improvement was due to some general biasing factor (such as teachers' reputations among students) or due to increased effectiveness of the teachers because of added experience is not clear.

Influence of Student Factors on Student Ratings of Instructor Effectiveness

As in the case of teacher factors, the studies concerned with student factors other than grades have been sporadic and not too clearly defined. Often they are just a by-product of studies concerned with other aspects of student ratings. Available studies have been summarized in Table 11. Information on four factors was considered: size of class, sex of students, age or maturity of students, and intelligence or mental age of students. By and large the results of the various studies show that these factors have little bearing on student rating. The curvilinear results found by Starrak in regard to influence of size of class are of some interest. It is unfortunate that Heilman and Armentrout did not test for curvilinearity as the size of the classes in their study ranged from 17 to 121. Starrak concluded: "On the basis of the ratings, 20 students seem to be the optimum number for a college class." Although his study was extensive (ratings were made quarterly on all instructors of the college and cover several years with a total of 40,000 ratings), it is difficult to see how the optimum size of a class could be selected merely on the basis of student ratings.

In the case of the influence of the sex of the pupils it might well be expected that girls and boys would differ in their ratings of teachers of certain subject matter. It is possible that a woman teacher better understands the emotions and thinking of girl students while a man teacher might deal better with boys and that these differences might vary for different student age groups. To a limited extent the few studies on this

Table 11

Relationship of Student Factors to Student Rating

Investigator	Teacher sample	Number students per teacher	Type of rating	Student factor	Correlation
Guthrie (1927)	87 college	16.5	Graphic	New students vs. advanced students	Ranking by new students had lower reliability
Remmers (1929)	115 college	74.8	Purdue scale	Size of class	Small classes (less than 10 pupils) tend to give higher ratings; otherwise no relationship
Bowman (1934)	21 student	10-42	General merit	Boys vs. girls	.64
Starrak (1934)	Entire college faculty	(Not reported)	Graphic (17 items)	Size of class	Slight relationship--classes with less than 7 and more than 50 tended to give consistently lower ratings
	Entire college faculty	(Not reported)	Graphic (11 items)	Maturity	Little relationship
Heilman & Armentrout (1936)	46 college	17-121	Purdue scale	Size of class	.24
	46 college	42	Purdue scale	Jr. college vs. sr. college	Mean of sr. college higher but not significantly so
Bryan (1937)	22 sr. high school 41 jr. high school	20-152	Graphic (11 items)	Boys vs. girls	Significant differences for some teachers on varying number traits (19% of total number ratings) no differences for other teachers
	22 sr. high school	20-152	Graphic (11 items)	Maturity	Reliability coefficient ranged .69 to .91
	41 jr. high school	20-152	Graphic (11 items)	Maturity	Reliability coefficient ranged .61 to .94
	22 sr. high school	20-152	General merit	Mental age	-.19
	41 jr. high school	20-152	General merit	Mental age	.01
Ferguson & Hovde (1942)	1 high school	304	Graphic (12 items)	Maturity	Difference not significant
Davenport (1944)	51 high school	(Not reported)	Graphic (25 items)	Boys vs. girls	Girls tended to rate all teachers higher than did boys. Boys ranked male teachers significantly higher (.01 level) than female. Girls also rated male teachers higher (not significant).
Henriksen (1949)	(900 ratings)	(150 total)	Graphic (6 items)	Boys vs. girls	High agreement
Drucker & Remmers (1951)	92 college	1-8 251 students & 138 alumni	Purdue scale	Students vs. alumni	
				Sense of proportion & humor	CR 3.05 (students higher)
				Self-reliance and confidence	CR 3.05 (students higher)
				Fairness in grading	CR 2.32 (alumni higher)
				Interest in subject	CR 2.88 (students higher)
				(Rest of 10 traits not significant)	
	17 college	8		Sum 8 students vs. sum 8 alumni	.40 to .68 (10 items)
Bendig (1952)	2 college	23-24	Graphic (Miami Univ. Instructor Rating Sheet)	College class	(Upper classes more critical) Significant at .01 level.
	2 college	23-24	Graphic (Miami Univ. Instructor Rating Sheet)	Men vs. women	(Women the more critical) Significant at .01 level

variable appear to support these generalizations though the most outstanding result is the lack of differences between ratings by the two groups.

Investigations in which maturity or age of students was one of the variables studied appear to be unanimous in the conclusion that this factor influenced ratings very little. It should be pointed out though that in almost every case a very limited range in age of students was studied. Usually an investigation covered the range within a particular college or high school or was concerned with first year students as compared with advanced students regardless of age. The study by Drucker and Remmers (105) is an exception in that it dealt with the relationship between ratings by students and ratings by alumni of at least ten years' standing. This study is particularly relevant to the frequently raised objection to student ratings that students are too immature to rate their instructors and that many years later, as alumni, students will have different values and will evaluate their former instructors on a different and presumably better basis. Positive relationship of some magnitude was found. What differences did occur showed that the students ranked their instructors higher than did the alumni. The difference was significant for three traits. It is possible that this might reflect a change in the teachers, i.e., that they became more effective, rather than a change in opinion of students as they get older. There was high agreement between the students and alumni as to the relative importance of the ten traits on the scale. The Pearson product-moment correlation coefficient between median rankings of these ten traits by the 251 students and 138 alumni was .92.

Using Student Rating for Instructor Improvement

There appears to be considerable opinion that, properly used, student rating has value in bringing about instructor improvement. For example, Schutte (296), Clem (77), Flinn (116), Riley et al. (276), and Stuit and Ebel (327), after having students rate instructors on one form or another, state (generally without adequate research evidence) that student rating enables instructors to evaluate their courses and teaching performances and that students' opinions often provide a better basis for self-study and instructor self-improvement than do the opinions of supervisors.

At the end of both the first and second semesters Bryan (62), in 1938, asked pupils to rate 29 junior high school teachers. He used a 9-item, 5-point scale, defined in descriptive phrases. Improvement revealed by the ratings was reported in terms of the percentage of items showing a difference between the first and second ratings. In this and subsequent articles (63, 64, 65, 66) he indicated that most teachers find the student ratings helpful or, at least, not harmful. This expressed attitude of the teachers, however, may reflect a positive bias, in that participation of the teachers in the study was voluntary; thus, the population studied may have been one that already believed in the helpfulness of students' ratings.

In 1941 Ward et al. (348), using the Purdue Rating Scale for Instructors, asked students to rate 40 practice teachers at the end of one month of instruction and again at the end of the semester. The ratings were used in diagnosing the weaknesses of the practice teachers and as stimuli for improvement. On the retest 39 of the 40 teachers showed a gain in rating. Apparently no use was made of a control group of practice teachers who did not get information concerning themselves from student ratings against which changes in the experimental group could have been compared.

Porter (257), who based his opinion on a consideration of pupil ratings of 27 student teachers obtained in 1942, suggested that supervisors' ratings may be made more objective by making use of pupil ratings. Presumably Porter intended that supervisors should utilize pupil evaluation of practice teaching to support their own evaluation of practice teachers. Whether or not supervisory estimates thereby become more objective has not been established.

Self-Rating of Instructor Effectiveness

Few studies of self-appraisal by teachers have been reported in the literature. Surveys of rating practices in the schools also show that self-ratings are sparingly used.

In 1927 Roberts and Draper (279) reported results of a study of principals' reports obtained from 441 high schools with enrollments ranging from 5 to 4000 pupils in all sections of the United States. Of the 398 reporting on the use of self-ratings, principals indicated that in 86 schools teachers were required to rate themselves, in 3 schools it was suggested that they do so, and in 309 schools no such rating was required.

In 1945 Reavis and Cooper (262) surveyed 123 cities in 34 states and the District of Columbia. Only one of these required a report of self-appraisal filed for administrative evaluation.

Table 12 summarizes seven investigations. In six of these investigations, correlations were determined between self-ratings and certain other measures of effectiveness. Administrative ratings, pupil ratings, or pupil gain show negligible relationships with teachers' self-ratings. Seven of the 10 coefficients for different schools reported by Cooke (81) were .21 or less. Even the largest, an r of .94, is not significant, having been obtained with an N of only 25 teachers. The only coefficients significantly different from zero (at the .01 level) are those obtained by Flory (117) between self-ratings and ratings by friends. Unfortunately Flory did not report the difference between means of self-ratings and ratings of friends; hence, he provided no information pertinent to the question as to the tendency to overrate oneself. The close agreement between self-rating and principal's rating in the study by Fichandler (111) might be explained in part by the teacher's familiarity with the principal's former rating.

Table 12

Relationship of Self-Rating to Other Measures of Instructor Effectiveness

Investigator	Teacher sample	Measures compared	Correlation coefficient
Flory (1930)	35 student	Self-rating vs. composite rating 5 friends (same scale)	.56
	99 students	Self-rating vs. composite rating 2 friends (same scale)	.49
Ulman (1930)	116 with 1 sem. experience	Freyd's graphic self-rating scale vs. superintendent or principal's ratings	.09
Barr *et al.* (1935)	66 elementary	Torgerson self-rating scale vs. pupil gain (Stanford Achievement)	-.01
Cooke (1937)	9 to 48 elementary & high school	Self-rating vs. composite of 3 ratings by principal	.02 to .49
Davenport (1944)	44 sr. high school	Self-ratings ("How I Teach") vs. pupils ratings ("How Teachers Teach")	.25
		Self-ratings vs. pupils' ranking (how well teacher liked)	.02
Fuller (1946)	45 student	10-point graphic self-rating scale vs. composite supervisor rating	.09

The tendency for individuals to overrate themselves is exemplified in the study of Knight and Franzen (179) who, in 1922, asked 110 students to rate themselves in terms of order of interests and also to rate "ideal" and "typical" junior students. The correlation coefficients obtained between self-rating and rating for the ideal was .46 and between order of interests for ideal and typical students was -.64. The authors conclude that the data show a well-marked tendency for a person to overrate himself when he compares himself with others and that the tendency still persists when the judgment is independent of comparison with others.

In only rare instances are an individual's own estimates of his competence accepted at full value by his superiors. The educational field appears to be no exception in this respect. On the basis of the few available studies of self-ratings of instructors as well as from self-ratings in general, the obvious, undisguised self-rating scale technique would seem to offer little encouragement for further investigation. It is possible, however, that there may be some justification for further exploratory work with more subtle self-rating instruments.

Objective Observation of Instructor Performance

The emphasis of present day teacher-training institutions appears to be less upon selection of a particular kind of person than upon trying to teach methods of performance that will insure success in the classroom. The establishment of departments of instructor training at various Air Force bases attests to the adherence to this approach in the Air Force. Potential instructors are given training in methodology and provided with the opportunity to practice the approved techniques under simulated classroom conditions. In keeping with this emphasis upon instructor performance, it might be expected that an instructor's effectiveness might be evaluated by observing what the instructor actually does in the classroom, provided that the observed behaviors are validated against other criteria.

Investigations using observational methods to determine differences in performance of effective and ineffective teachers have been few in number and have varied widely in design. Brownell (59) points out this lack, stating that the use of the technique of continuous, or a series of spaced, observations intended to detect changes in some form of behavior has been grossly neglected in the research work in this area.

Unfortunately, also, most of these studies have leaned rather heavily upon the subjective judgment of the observers. In many cases the investigator himself, and sometimes an administrative official, did the observing though there are a number of studies in which specially trained independent observers have been employed. The observational methods used include chiefly variations of the time-sampling technique or check-list records of the presence, absence, or duration of particular activities. In a very few cases photographic, phonographic, stenographic reports, and frequency counts have also been utilized. Studies in which a rating scale was completed by an individual after observing a classroom situation are not included in this section.

POOR ORIGINAL COPY - BEST
AVAILABLE AT TIME FILMED

Reliability of Objective Observation

In only a few of the studies using the observational approach was the question of the reliability of the method considered. Too often it is thought enough to say that the observer has had practice in observing, or reliability was assumed on the basis of the fact that the observer was supposedly an "expert" in the educational field. These assumptions are made particularly, of course, in cases where the investigator or an administrator was the observer.

Where reliability was computed, the criterion most generally used was agreement of independent observers determined by use of a correlation coefficient or percentage of agreement on the basis of an item-by-item comparison of records. In a few cases occasion-to-occasion reliability was computed for the same observer. In Table 13 reliability coefficients are listed. In general, it may be said that the reliability of planned observational recording compares favorably with that of other methods. Anderson and Brewer (7) found that a total of from 300 to 400 minutes of observation yielded a high degree of consistency in the sampling of teachers' behavior and that observers were more reliable in recording "domination" than "integration."

Validity of Objective Observation

The most general criterion of validity of observation has been face validity. In a few studies, however, different methods of evaluating the same lessons were compared. In 1930 McAfee (208), who evaluated teacher efficiency by counting the number of good teaching practices and the number of poor practices as recorded by one observer on a detailed rating sheet, obtained a correlation coefficient of .41 between this evaluation and supervisory ratings for a group of 98 teachers. Shannon (304), in 1936, compared three methods for measuring efficiency in teaching. One of these was based on an attention score obtained by dividing total minutes of observed pupil attention (determined by pupil's postural attitudes and movements) by total possible minutes of pupil attention. The other two, which were subjective, although accomplished by the same individuals as the attention score, consisted of five-point ratings made on a score card containing 43 rubrics grouped under five headings, and ranking of the teaching performance of each teacher within his group. The observer-raters were 14 graduate students who had had experience in supervision, and the teachers studied were 111 student teachers divided into eight homogeneous groups. Correlations between score-card ratings and attention scores ranged from .07 to .61 and between rankings and attention scores from -.16 to .73 while the correlations between score-card ratings and ranking ranged from .38 to .97. It appears that while pupils' attention scores are more reliable (see Table 13) than the score-card ratings or ranking they do not compare as closely with the ratings or ranking as the latter two

Table 13

Reliability of Various Methods of Observing Teaching Effectiveness

Investigator	Teacher sample	Observational Method	Measure of reliability	Reliability coefficient
Wrightstone (1935)	(Not reported)	Recording activities by use of categories or codes of behaviors	3 observers vs. 3 observers	.91 (initiative)
			3 observers vs. 3 observers	.93 (cooperation)
			One observer (odd-even days) for 5 classes	.82 to .96
Shannon (1936)	111 students	Attention scores	2 observers vs. 2 observers	.67 to .84
		Rating on score card	2 observers vs. 2 observers	.49 to .84
		Ranking	2 observers vs. 2 observers	.26 to .92
Jayne (1945)	28 elementary rural	Specific observable activities (check list)	Scores from 2 observations	.41 to .95
	10 elementary rural			.66 to .99
Anderson *et al.* (1945)	3 elementary	Time-sampling of dominative and integrative behavior	Observer vs. observer	.85 to .93 (domination) (Per cent agreement 80 to 86) .74 to .91 (integrative) (Per cent agreement 74 to 78)
Ryans (1952)	48 elementary	Classroom observation scale (check list of 26 behavior dimensions)	Scores from 2 observations for 4 observers	.72 to .83
		Over-all assessment	Intercorrelation between observers	.68 to .84

compare with each other. Since the rankings determined by the two subjective measures bear higher correlations than comparisons involving attention scores, Shannon concludes (gratuitously, it appears to the reviewers) that "the more subjective means are the better ones of the three included in this investigation."

In a later paper in 1942, Shannon (307) made another study of the validity of attention scores. Two seventh and eighth grade classes composed of 47 boys and 53 girls were used. Observations were made by three graduate students while material was read to the class. Pupils were later given multiple-choice tests covering the material read. Correlations between attention scores and test scores were: for boys, .67; for girls, .34; for total group, .59. The respective correlations between test scores and intelligence were .37, .40, and .37, while attention and intelligence correlated .14, .34, and .21. The author concluded, "Assuming that the material read...was new to the children the evidence is damaging to the validity of the attention measurement. That it has a slight degree of validity is clear, but that it has enough validity to warrant its use in judging classroom activity is worse than doubtful." It appears to the reviewers that Shannon was unduly pessimistic. Results showing an attention measure which is somewhat more closely related to student performance than it is to intelligence have implications justifying further research. Strictly speaking, Shannon's study does not pertain to the teaching but rather to pupil factors effecting learning, since the teaching was the same for all pupils.

Some Significant Observational Studies

The findings of a number of studies using the observational method will be reviewed at some length because the results appear distinctly encouraging.

One of the earlier observation studies was that of Barr (16), in 1929, who set forth to observe characteristic differences in teaching performance of good and poor teachers of the social studies. A group of 47 superior teachers was selected, on the basis of superintendents' and state inspectors' ratings, from cities with a population of 4000 and over. Similarly, 47 poor teachers were selected from cities of less than 4000, excluding teachers from one- and two-room rural schools. The superior teachers were from the "promoted" group, with better training and more experience than the poor teachers. The poor teachers were rated C- or below, and 50% did not return to their teaching positions the following year. The median experience of the good teachers was 12.3 years, while that of the poor teachers was 3.7 years. An obvious defect of the design of this study was the failure to hold teaching situation constant by holding type of school constant.

Teaching methods were studied by using a combination of subjective and objective devices. These included: (1) an annotated stenographic report, (2) a time-chart record of one or more recitations, (3) an attention chart for one or more recitations, (4) a time-distribution study of the major activities of the recitation periods for one week, (5) a check-list record of one recitation, (6) a comprehensive questionnaire upon the various practices of each teacher, (7) superintendents' estimates of the teachers' strengths and weaknesses, (8) the teacher's self-analysis of her teaching.

Barr found the usual subjectively determined qualitative differences between good and poor teachers. Strong points of superior social study teachers included, for instance, knowledge of subject matter, good technique in asking questions, ability to stimulate interest, and socialization of class work. Elements of weakness included such items as no provision for individual differences, formal textbook teaching, no interest in work, no daily preparation, weak discipline, and no knowledge of subject matter. Barr mentioned 52 separate traits in listing the personal qualities of good and poor teachers, including personal appearance, sincerity, energy and vitality, and speaking voice. Barr's results may be somewhat suspect since his evaluation of the qualitative differences may have been unintentionally contaminated by foreknowledge of the identity of the good and poor teachers. With respect to quantitative differences he found that correlations between time distributions of various aspects of class activities and supervisory ratings ranged from -.23 to .17. Relationships between particular items on the time-chart record and estimates of teaching success were also found to be small. Barr concludes that it is doubtful "whether time expended in class upon such items as those reported in this study are reliable indices of teaching ability." He indicates that within very broad limits there appear to be no optimum time expenditures for class activities and that good teachers function successfully within a wide range of time expenditures.

Olson and Wilkinson (248), in 1938, attempted to investigate teacher personality as revealed by the amount and kind of verbal direction used in behavioral control. They used time-sampling records of responses of 30 student teachers, 25 women and 5 men, to a constant group of children, 13 first grade, 13 third grade, and 13 fifth grade pupils, in a one-room situation. Each of these grade groups was divided into two subgroups or classes, equated as nearly as possible for ability. Each teacher was observed with each one of the subgroups at least once. Ten five-minute samples per teacher were obtained for each class taught. The frequency and methods of redirecting children's attention were observed. Distinction was made between language and gestural responses and between positive, directive verbal responses as opposed to negative responses. A "blanket score" was also obtained by noting each five-minute period in which the teacher adjusted to the class as a whole, rather than to an individual in controlling behavior when the attention of an individual child needed to be redirected. Observations were made by a critic teacher. Teacher efficiency was obtained for each grade, based on independent judgments of school principal and critic teacher together with average ratings obtained on Leonard's Rating Sheet for Predicting Teaching Success. The coefficient of correlation between the two raters was .73 for the total score on the scale. The correlation between rated teacher efficiency and total teacher response score was -.06, between teacher efficiency and positive teacher response, .59, and between teacher efficiency and blanket response, -.62. When correlations were computed between teacher responses of the five most able teachers and pupils' scores on the Haggerty-Olson-Wickman Behavior Rating Scale, Schedule B, (pupils were rated by principal and critic teacher) the resulting coefficient was .69. For the five least able the coefficient was .30. Olson and Wilkinson felt their results indicated that there was better distribution of attention in terms of pupil need in the case of the able teachers and a quantitative analysis showed that the less able teachers tended to avoid contact with the more difficult cases. Conclusions based on correlations involving two N's of five each, however, cannot be taken too seriously.

Jayne (166), in 1945, compared pupil changes with specific observable teacher activities. He used 28 teachers of Rostker's (282) study, and an additional 10 teachers and 95 pupils. Pupil gain for the 28 teachers was measured by computing residual gain (actual gain minus predicted gain) for classes in social studies on the basis of eight tests, six of which were published tests and two composed for the particular course of study. For the 10 additional teachers, gain was measured after each class had had a lesson on Alaska by computing posttest minus pretest and recall test minus pretest. In this study no single, specific observable teacher act was found whose frequency or per cent of occurrence was invariably significantly correlated with pupil gain. "There is," Jayne states, "in general, little relationship between specific observable teacher acts and the pupil-gain criterion." The results, however, varied greatly for different methods of assessing pupil gain.

Jayne noted that analysis of the coefficients of correlation seemed to indicate that the most significant positive correlations with pupil gain were those having to do with extent to which questions were based on pupil interest and experience rather than on assigned text, the extent to which the teacher challenged pupils to support ideas, and amount of spontaneous pupil discussion. A composite index score, called "Index of Meaningful Discussion," based on seven items, correlated .80 with pupil gain based on a composite of eight tests and .39 with pupil gain based on two tests constructed for the particular course for the 28 teachers from the Rostker study; however, this score yielded negative coefficients of -.67 for immediate recall and -.68 for delayed recall for the 10 additional teachers. Jayne explains this by the fact that the aim of the lessons in the first study (Rostker's) and the second were different. The teaching in the first study was of wider scope, while that of the second was aimed toward recall, making discussion of textbook material essential. Accordingly, Jayne made up a second composite of items relating to mere recall of assigned material. This yielded higher coefficients for the group of 10 teachers (.82 for immediate recall and .53 for delayed recall) than it did for the 28 teachers (.19 for the composite of eight tests and -.35 for the course tests). From this it would seem that teaching procedures that were appropriate and effective under conditions of the first study may have been inappropriate and ineffective under conditions of the second study.

Anderson, Brewer, and Reed have made a series of rather exhaustive studies of teachers' classroom behavior. In the first of their studies in 1945 Anderson and Brewer (6) investigated dominative and socially integrative behavior of kindergarten teachers. A total of 101 children in two schools were observed to determine pupil reaction to the differential behavior of teachers. Among other results, teachers were found to use domination of individual children more consistently than integrative contacts; teachers tended to dominate boys more often than girls; the number of teacher-pupil contacts per hour had little relation to the numbers of children in the room; for a mental hygiene point of view, there was "better" teaching in the morning than in the afternoon. It thus appears that individual children may live in vastly different psychological environments in the same schoolroom.

In a subsequent monograph in 1946 Anderson and Brewer (7) discussed results of observations of teachers' dominative and integrative contacts in second, fourth, and sixth grades. The categories of teacher behavior observed were largely descriptive and represented activities that made a difference in the behavior of the children. Fourteen statistically significant differences between children in the two second grade classrooms were found. These were reported to be consistent with the personality differences of the teachers. Pupils of the more integrative teacher showed significantly lower frequencies of looking up, playing with foreign objects, in general less conforming and nonconforming behavior, and more spontaneity, initiative, and social behavior than did those of the dominative teacher. Teacher contacts in the sixth grade situation were as frequent as they were in the second and fourth grades.

In a third monograph in 1946 Anderson, Brewer, and Reed (8) report on follow-up studies of the effects of dominative and integrative contacts on children's behavior. The dominating teacher was, a year later, still dominating, but the children who had passed on into the third grade no longer showed the undesirable personality patterns formerly noted. Two third grade teachers were also observed, one of whom had twice as many frequencies of domination in conflict contacts with individual children and over four times as many such contacts with groups of children as the other teacher. Within the validity of certain mental hygiene assumptions, observations of the teachers' classroom behavior revealed certain strong points and certain weak points. The authors suggest that the weak points are such that they are amenable to correction by instituting teacher inservice training programs. As a result of the work by Anderson _et al_.

discussed in the above references, a scale for recording dominative and integrative behaviors of teachers has been prepared and is to be published in a forthcoming issue of the Applied Psychology Monographs.

In 1952 Ryans (288, 289) reported two studies concerned with factor analysis of teacher behaviors, one of elementary women teachers (275 third and fourth grade) and one of high school teachers (115 men and 134 women). These investigations are part of the "Teacher Characteristic Study" being conducted by the American Council on Education and the Grant Foundation. The purposes of this broader project as outlined are "(1) to try to determine the personality patterns of teachers,(at elementary and secondary school levels) and 2) to explore the possibility of developing measures that will reflect, and predict, such patterns as may be found." The research is limited to the study of the personal qualities of the teacher on the assumption that certain minima of intelligence and knowledge of subject matter (and perhaps knowledge of "techniques" of teaching) are primary requisites for teaching. In the part reported by Ryans, observers trained over a period of five weeks recorded observations on a specially devised Classroom Observation Scale. This scale covered 26 behavior dimensions relating directly to teacher behavior and pupil behavior (presumably reflecting teacher behavior). Each of these dimensions of behavior was described in terms of opposite poles and was assessed on a four-point scale. Each elementary teacher was observed by at least three different observers on different occasions. Each high school teacher was observed by at least two different observers and sometimes by three. Data were factor analyzed by the centroid method. The factors obtained for the two groups of teachers did not duplicate each other entirely although there are points of similarity. Ryans (287) believes that three correlated factors may serve satisfactorily to describe teacher behavior at both levels: (1) understanding, friendliness, and responsiveness on the part of the teacher; (2) systematic and responsible teacher behavior; and (3) the teacher's stimulating and original behavior. The three factors show somewhat different relationships in the two school situations. Factors 1 and 3 are most highly correlated in the elementary school situation with Factor 2 being relatively independent. In the secondary school situation Factors 2 and 3 are most highly related with Factor 1 being relatively independent.

The work reviewed in the foregoing section constitutes a preliminary attack which promises to be one of the most productive in this area. Systematic observation should prove fruitful both as a source of rationale hypotheses concerning the nature of teacher effectiveness and as a technique for testing such hypotheses. The relevant categories for observation will of course depend on the particular situation being investigated. Thus, in Air Force schools, for instance, the observational technique will probably employ categories which differ from the categories of observation developed for elementary and secondary school teacher behavior. The differentiation of those behavior categories which are related to instructor effectiveness from those which are immaterial remains to be investigated.

Another approach to the investigation of the effectiveness of instructors should also be explored further. It is that in which teacher factors, situation, or method are systematically varied as was done, for example, in studies (204, 355) of so-called authoritarian-democratic teaching. It has been suggested that the experimental classroom in which factors associated with teaching can be manipulated under controlled conditions may offer greater potentialities for achieving successful results than do the correlational studies of teaching competence in situ.

Student Change as a Measure of Instructor Effectiveness

Most educational authorities hold that the primary responsibility of the instructor is to bring about change in the knowledge, skills, understandings, attitudes, appreciations, interest, and motivation of his students. For advocates of this point of view the determination of instructor effectiveness is logical and straightforward. It consists of measuring the changes that are produced in students as a result of the instructor's efforts.

The importance of pupil achievement as a measure of teaching ability has long been recognized. As early as 1921 Courtis (85) pointed out the significance of student gains as a criterion of teaching efficiency, as well as the importance of holding constant extraneous factors. He pointed out that a comparison of pupils' learning curves for incidental learning with their curves for direct instruction would provide a means of evaluating teacher competence. In a later article (86) he cautioned that any method of measuring teaching effectiveness must involve the use of a "single-variable" measure. He held that it was necessary to measure the change in the rate of growth which takes place in the student when the amount of quality of teaching is the only variable in which change occurs. Courtis then defined good or poor teaching by the periods when the actual growth curve showed marked deviation from the theoretical growth curve. To illustrate the method, an observed growth curve of a particular function was compared with a theoretical growth curve for the same function as defined by Gompertz's formula expressing the general law of biologic growth. The author maintained that, while much research remained to be done, an exact scientific method had been devised by which the effects of teaching might be precisely measured.

Unfortunately the possibility of comparing curves of "incidental learning" with curves of learning from "direct instruction" seems much further away today than it did to Courtis in 1921. While there has been immense progress in the science of measurement, this progress has brought a realization of the difficulties involved in charting intellectual growth curves, particularly in an area as ill-defined as "incidental learning."

The first reported attempt to use student change as a measure of instructor effectiveness appears to have been that of Hill (155) in 1921.

This and subsequent studies can, for purposes of discussion, be divided into five classes, according to the kind of measure of student change that was used or suggested: raw gain (posttest minus pretest scores); achievement or accomplishment quotient; miscellaneous measures; corrected raw gain (raw gain corrected for initial intelligence, grade, or other variable); and residual gain (actual gain minus predicted gain).

Among the investigators using raw gain as their criterion or as one of their criteria are Baird and Bates (14), Barr et al. (20), Betts (33), Bimson (34), Bowden (46), Brookover (55), Hartmann (146), and Hill (155). Use of raw gain as a criterion is manifestly inadequate. Teaching is only one among many factors operating to produce changes in students. It is necessary, consequently, to hold constant all factors other than the effects of the particular teaching situation being studied. Since the early 1930's raw gain has rarely been used or, if used, was one of several gain criteria.

The accomplishment quotient or ratio which is the ratio a pupil's educational age or quotient, as measured by standardized achievement tests, bears to his mental age or quotient, as measured by standardized intelligence tests, has been widely used as a so-called objective measure of teaching efficiency. This ratio allegedly indicates the extent to which a child is "working up to his ability." Goodenough (129) points out, however, that there are several sources of error which are likely to reinforce rather than cancel each other both for individual cases and in group data. The errors arise from lack of knowledge as to the absolute zero point in the two measures, from unequal variability, and from failure to allow for regression due to errors of measurement. As Goodenough (129) says "...in spite of repeated demonstrations of the unsound assumption upon which the method is based, it has proved to be one of the most persistent die-hards in the history of educational psychology." The accomplishment or achievement quotient has been used by Barr et al. (20), Coy (88), Crabbs (89), Simmons (310), and Stephens and Lichtenstein (323).

Certain investigators have attempted to use other student measures as criteria of instructor effectiveness. Thus, in 1934 Davis (96) used pupil achievement in terms of passing or failing state high school examinations; in 1934 Frederick and Hollister (121) used numbers of honor grades and failing grades; in 1935 Lancelot (192) utilized persistence in taking advanced courses and grades received in those courses; in 1938 Beaumont (26) employed number and achievement of students taking advanced courses; in 1945 Cheydleur (75) used ranking of instructors according to the ratio of class average to group average in college French. While some differences among instructors were found, the outcome of none of these studies appeared to be very significant.

The validity of these student measures as criteria of instructor effectiveness may well be questioned. Whether or not a given student passes or fails a state examination, or achieves honor or failing grades, depends upon many factors besides his teacher. The same is true of the ratio a class average bears to a group average. Where pupils from different schools are compared, some means must be found for controlling such variables as size and type of school, equipment and library facilities, and the like. In all cases where groups or classes of pupils are compared, such pupil factors as intelligence, motivation, interest, and aptitude of pupil for a particular subject must also be controlled. The reliability and validity of the examinations on which the student's grade is based must also be taken into consideration. The number of students or their persistence in taking advanced courses and the grades achieved in these courses may depend upon the enthusiasm of the instructor or the interest he is able to build up in his students in the elementary courses or it may be a function of the reputation or competence of instructors teaching the advanced courses. Any simple measure of student gains that fails to take into account the complexities involved will almost inevitably produce misleading results.

While not strictly concerned with gains Seyfert and Tyndal (302), in 1934, used a rather unique approach in attempting to evaluate differences in teaching ability. The subjects were two general science teachers who had previously been rated best and poorest of a group of seven teachers by superintendent, principal, and supervisors. Four groups of students were used: two groups of girls matched for age and score on the Terman Intelligence Test and two mixed groups with age and score on the Rulon Science Test held constant. Student achievement was determined in terms of the mental age necessary in order that a student of the less able of two teachers may achieve the same score level as a corresponding student of a better teacher. The difference in teaching ability between the two teachers was found to be equivalent to about three months of mental growth on the part of the students.

Lancelot (191) says that mere acquisition tests are not sufficient to determine student gains because of the discrepancy between acquisition of knowledge on the one hand and its retention on the other. He feels that a better and relatively sound criterion of teaching ability consists in the degree of retention by the students of knowledge taught. While theoretically this may be true, use of amount of retention as a criterion poses the additional problem of finding some method for holding intervening learning constant.

The first studies to measure student gains by partialling out factors other than achievement were those of Moss *et al.* (235) in 1929, Taylor (331) in 1930, and Betts (32) in 1933. Moss *et al.* in studying the efficiency of chemistry instructors used classes equated for intelligence and previous training in chemistry. Taylor corrected for intial score, age, and intelligence. Betts, besides using a measure of gain in reading indicated by the mean of the final scores on the Stanford Achievement Test, studied the relationship of various teacher measures with standard deviation of the class and measures of heterogeneity and homogeneity of achievement which were obtained by combining pupil mean final score and standard deviation by formulas. He also computed correlations with these teacher measures after partialling out factors of age, initial score, and standard deviation. He obtained much higher correlations for his teacher measures (intelligence, professional information, vocabulary) when the criterion of "hetero-achievement" was used. He points out the pitfalls of judging gain by score alone or by heterogeneity (standard deviation) of the group. The latter "can be secured by causing dull pupils to forget some of the things they knew initially and by inducing superior pupils to learn. If both average achievement and heterogeneity of pupil groups are taken in combination, such an influence serves to reduce the composite score because a maximum composite can be obtained only by increasing both concurrently."

In 1945 Bolton (41) used the ratio of mean pupil achievement to its standard deviation as a measure of teaching effectiveness. In comparing six teachers of United States History for matched groups of pupils, he reported that one teacher excelled, having a ratio of teaching effectiveness more than four times greater than the teacher next in line, while the ratios of the other five were close together. In interpreting Bolton's findings one should avoid the fallacy of the tobacco company that advertises cigarettes which contain "five times less acid tar." The use of ratios based on educational or psychological test scores involves assumptions untrue of such scores, namely that their lower limit represents an absolute zero point and that intervals between scores are equal. We can never say that one person is four times as intelligent, knows twice as much history, or is four times more effective as a teacher than some other person. Other investigators who have used corrected raw gain included Bimson (34), Day (98), and Georges (127).

Of the several methods used to measure pupil change, residual pupil gain (i.e., the difference between actual gain and predicted gain) is becoming more widely used as a criterion of instructor effectiveness. This method is really a more refined example of the corrected raw gain criterion already discussed. Its main advantage is that a more adequate attempt is made to hold constant student factors other than the effect of the instructor. The chief disadvantages are its dependence upon the availability of valid instruments for measuring student growth, the excessive time required to obtain the necessary data, and the rather elaborate statistical assumptions and analysis involved. With all its difficulties, however, this appears to be one of the best criteria of instructor effectiveness.

Several versions of residual pupil gain where gain was predicted on the bases of such student factors as initial scores or intelligence quotients have been used by Gotham (132), Jayne (166), Jones (172), LaDuke (188), Lins (203), Remmers et al. (269), Riesch (275), Rolfe (280), Rostker (282), and Von Haden (344). These studies will be considered on subsequent pages.

Difficulties of the Gains Criterion

Tyler (338) and others, however, have pointed out the difficulties which attend the use of student gains as a criterion. In the first place, as was noted earlier, what is meant by gain must be adequately defined. An instructor is called upon to perform many duties and to accomplish many changes in his students that are not measurable in terms of subject-matter achievement. Therefore, any measure or measures of student change based on gain in subject matter alone represents only a small area of the instructor's total effectiveness. This objection probably applies less or may not be applicable at all to the Air Force situation, in which the instructor's chief concern is the teaching of course material of a technical nature.

Determination of gains attributable solely to the teacher is dependent on the availability of valid instruments for measuring such growth. If more than just subject-matter learning is to be used, mere use of achievement test data is manifestly inadequate.

As a practical solution most studies measure student gain on the basis of subject matter learned on the assumption that it is, if not the total gain, at least probably representative of the major part of the teacher's job. Even assuming that the type of gain that is to be measured is known, there are still difficulties in obtaining a valid measure. If gains of classes under different schools are compared, use of standardized achievement tests may only reflect the differences in the teaching program in use in the different schools and not the ability of the different teachers. In this connection, tests designed to measure the learning achieved in a given course of study are probably more adequate than the more general standardized achievement tests. The nature of the subject matter selected may also make a difference. A gain in spelling may be a less complex measure than a gain in arithmetic. Judging the effectiveness of a teacher who is teaching several subjects, such as is usual in the elementary grades, on the basis of the gain of his students in a single subject field is obviously inadequate.

As another difficulty, an instructor whose students obtained high initial scores might show up poorly under a gains measure even if correction were made for the high scores. This is because of the limited gain possible in the case of high original scores and the increased improbability of making a given gain as the initial score becomes higher. Every test has a ceiling, a maximum or perfect score beyond which no one can go. If a student's score is near the top on the initial test he cannot gain as much as the person whose score falls near the bottom. This difficulty can be overcome if regression equations are used to obtain predicted final scores and if the tests used have high enough ceilings. Analysis of covariance may also counteract this difficulty.

The gain of a student with a high initial score for his grade group is also limited to some extent by the general teaching situation. In most schools for each subject and each grade there is a definite range of difficulty of material to be taught. This in effect imposes a test ceiling for that particular grade in terms of the subject content considered to fall within its range. For this reason a student who has already made inroads into the subject-matter content for his grade will appear to be making less progress than a student of lower initial achievement.

Reliability of Student Gain

In Table 14 appear reliability coefficients of measures of pupil gain as reported by five investigators. It will be noted that, as compared with commonly reported test reliabilities, most of the coefficients appear to be rather low. Taylor (331) explained the reliability coefficient of .26 for reading progress in terms of the slight numerical changes in scores that took place. Rolfe (280) reported a reliability coefficient of .82 for the initial composite of three Hill tests and a coefficient of .78 for the final Hill composite, yet the reliability of the change was only .19. In general, reliabilities for gain tended to be lower than those reported for either initial or final scores. Rostker (282) suggested that this may have been due to the fact that the gain reliability coefficients contain errors of measurement derived from both the initial and final applications of the tests used. In addition, the reliability of a gains measure is dependent not only on the reliability of initial and final measures but also on the correlation between them. The higher the correlation of these variables the lower the reliability of the gains measure.

In general, the statistical computations involved in the estimation of the reliability of student gains are equivalent to those involved in estimating the reliability of differences between test scores. Methods are discussed and relevant formulas are given, for example, in Lindquist (202).

Interpretation of a reliability coefficient rests on the assumption that it has been obtained as the result of correlating comparable measures of the same thing and that the variable errors are uncorrelated with themselves and with the true scores. If errors are correlated, it follows that the obtained reliability coefficient will be spuriously high. In this connection it should be noted that all the correlations reported in Table 14 are split-half. These coefficients show the uniformity of the effect of the instructor within a single class. They do not give any information as to the consistency of instructor effectiveness in different classes. Coefficients of reliability obtained by the split-half method will be increased by any noninstructor variables that affect a whole class, while class-to-class correlations would be decreased by such influences.

An investigator may be interested in the effect of the instructor upon a class as a whole or upon certain types of students within a class. Since most research in this area has been concerned with the effectiveness of the instructor with respect to a class, measures used in determining pupil gain have usually consisted of means for groups of students. The reliability of average measures of pupil gain based on a group of pupils may differ from reliability of gain determined for individual pupils.

Table 14

Reliability of Measures of Student Gain

Investigator	Number of classes	Pupil measures	Measure of gain	Reliability[a] coefficient
Taylor (1930)	105 elementary classes	Arithmetic (Woody-McCall)	Achievement quotient	.76
		Reading (Thorndike)	Achievement quotient	.26
LaDuke (1945)	31 elementary classes (Community living)	Appreciation[b]	Residual gain	.64
		Attitude	Residual gain	.62
		Information	Residual gain	.72
		Interest	Residual gain	.53
Rolfe (1945)	47 elementary classes (Citizenship)	Unit[b] Composite (2 tests)	Residual gain	.46
		Wrightstone Composite (3 tests)	Residual gain	.69
		Hill Composite (3 tests)	Residual gain	.19
		UWH (Composite all 8 tests)	Residual gain	.68
Rostker (1945)	28 elementary classes (Social study)	Unit[b] Composite (2 tests)	Residual gain	.37
		Wrightstone Composite (3 tests)	Residual gain	.75
		Hill Composite (3 tests)	Residual gain	.44
Stephens & Lichtenstein (1947)	86 elementary classes	Arithmetic (Stanford Achievement)	Modified achievement quotient	.63

---

[a] Uncorrected, split-half of each class or of samples therefrom.

[b] Tests specially constructed for the particular course of study.

Correlation of Student Gain with Other Measures of Instructor Effectiveness

Investigations in which attempts have been made to relate measures of student gain to other presumed measures of instructor effectiveness have been summarized in Table 15. Reported coefficients range from -.61 to .81. In more than half of these studies one or more negative correlation coefficients were obtained. This extreme variability may mean that measures used were inadequate or that the gains criterion is dependent on factors other than the teacher such as subject matter taught or pupils' academic level. On the other hand, in view of the statistical pitfalls awaiting an unwary user of the student gains criterion, certain of the studies which show low or negative relationships may merely be reflecting inadequate research design.

In five of these studies, Simmons (310), Bimson (34), Brookover (55), Von Haden (344), Remmers et al. (269), correlation coefficients were not computed, were not significant, or were not available to the reviewers. Bimson consistently found that pupils of teachers rated above the median made higher gains than pupils of lower rated teachers, but that greater progress was made by pupils of lowest intelligence. It should be pointed out that Bimson (34) determined a progress quotient by dividing the difference between pretest and posttest scores by I.Q. This procedure appears highly questionable since it penalizes the brighter students who tend to make high initial scores. Due to test ceiling the possible gains of these students are less than possible gains of duller students. This in turn favors the instructor whose efforts are directed toward the students of low I.Q. The "little relationship" reported in Table 15 for Jayne's study (166) is based on the fact that Jayne found significant only 20 or about six per cent of 336 correlations between frequency scores of observable instructor activity and pupil gains. In the report reviewed, Brookover (55) failed to include statistical analyses which were evidently made in the original doctoral dissertation from which the article was drawn. The negative association which Brookover found between mean gains in pupils' history information and the pleasurable personal-relationship which the teacher has with his pupils is what might be expected. The instructor who spends his time being a "good fellow" with the students probably to some extent neglects to impart subject matter information.

As one examines the results of correlational studies such as some of those summarized in Table 15, one wonders what thinking lay behind the investigations. Some of the variables intercorrelated are so unreasonable and arbitrary that one suspects they were computed simply because data on certain variables were available or could be readily obtained. In some instances, certainly, there exist no psychological nor educational grounds on which relationship between student gain and some of the variables used might reasonably be expected to exist. Computation of such correlations were obviously largely a waste of time and their reporting makes no contribution to our understanding of the relationship of student gains to rated effectiveness of instructors.

Table 15

Correlation of Measures of Student Gain with Other Measures of Teacher Effectiveness

Investigator	Teacher sample	Subject	Measure of pupil gain	Measure of teacher effectiveness	Correlation
Hill (1921)	135 elementary	Arithmetic, penmanship, spelling	Posttest minus pretest	Administrator rating (Winnetka)	.45
				Administrator rating (Gary)	.24
				Administrator rating (Detroit)	.19
Crabbe (1925)	Elementary, rural	Reading	Achievement quotient	Average ranking (3 supervisors)	.27
	Elementary, urban	Reading	Achievement quotient	Ranking (1 supervisor)	-.36
	Elementary, rural	Composite 5 subjects (reading, arithmetic, spelling, penmanship, composition)	Achievement quotient	Estimating teaching in general	.32
	Elementary, urban		Achievement quotient	Estimating teaching in general	-.26
Baird & Bates (1929)	470 elementary	Reading	Achievement quotient	Principal rating (general merit)	.14
Taylor (1930)	105 elementary	Reading	Posttest minus pretest	Composite administrator ranking & education specialty rating	.24
		Arithmetic			.02
		Reading	(Initial scores, age & intelligence held constant)	Composite administrator ranking & education specialty rating	.24
					.10
Simmons (1932)	40 elementary	(Not reported)	Achievement quotient	3 administrator ratings	Negligible relation
Barr, et al. (1935)	66 elementary	Arithmetic	Posttest minus pretest	Superintendent rating (composite, 7 scales)	.09
		Arithmetic	Achievement quotient	Superintendent rating (composite, 7 scales)	-.04
		Arithmetic	Achievement quotient	Superintendent rating each of 7 scales	-.13 to .18
Jones (1946)	13 high school	English	Residual gain	Supervisor rating	-.61
	65 high school	15 high school subjects	Residual gain	Supervisor rating	-.28 & .10
Brandt (1949)	9 (LaDuke)	Community living	Residual gain (original study)	Supervisor follow-up (8 yr.)	.14
	17 (Rostker)	Social studies	Residual gain (original study)	Supervisor follow-up (11 yr.)	.35
Remmers, et al. (1949)	53 chemistry laboratory (28 exceed prediction; 25 under prediction)	Chemistry	Residual gain	Student rating 12 traits:	
				Care of communal apparatus	.01[b]
				Rating compared with Purdue instructor test	.01
				Knowledge of chemistry	.02
				Returning dailies & tests	.02
				Should instructor be kept	.02
				Supervision during tests	.02
				Coverage of assigned work	.02
	50 chemistry (28 exceed prediction; 20 under prediction)				
			Other traits	Not significant	
Von Haden (1945)	17 high school women, 1 yr. experience	6 high school subjects	Residual gain	Supervisor ratings of personal data items	None of 34 r's significant
Lins (1946)	17 high school women, 1 yr. experience	6 high school subjects	Residual gain	Composite 5 supervisor ratings	.19
				Pupil evaluation of teacher effectiveness	.06
Bimson (1937)	25 high school	Algebra, general science, history	Achievement quotient	Supervisor ratings	Higher rated teachers show consistently more pupil progress
Brookover (1945)	66 high school male	U. S. History information	Posttest minus pretest	Pupils' pleasant personal relations	Low significance negative relationship
	66 high school male	U. S. History	Posttest minus pretest	Administrator ratings	No significant relationship
	66 high school male	U. S. History	Posttest minus pretest	Pupil rating of ability	Low, irregular relationship
Gotham (1945)	57 elementary, 1- & 2-room	Citizenship course	Residual gain	Superintendent, supervisor, observer (3 scales)	.40
Jayne (1945)	28 elementary rural	Social studies	Residual gain	Frequency of observable activities	Little relationship between specific observable acts & pupil gain

---

[a] Exclusive of 1- and 2-room schools

[b] Level of confidence of difference in mean ratings between instructors whose classes obtained grades in chemistry higher than predicted and those whose classes obtained grades lower than predicted.

[c] Tests used: Townsend & Willis Coopérative Social Studies Test
Washburne Social Adjustment Inventory
Wood Right Conduct Test
Remmers' Scale for Measuring Attitude Toward Teacher

Table 15 (Cont.)

Investigator	Teacher sample	Subject	Measure of pupil gain	Measure of teacher effectiveness	Correlation
LaDuke (1945)	31 elementary, 1-room	Community living course	Residual gain-information test	Superintendent rating (3 scales) Supervisor teacher rating (3 scales)	.17 -.03
			Residual gain (Comprehension appreciation, attitudes, information, interest)	Superintendent rating (3 scales) Supervisor teacher rating (3 scales)	.02 -.25
Rostker (1945)	28 elementary, rural	Social studies	Residual gain	Investigator rating (3 scales)	.23, .26, .34
		Social studies	Residual gain	Supervisor rating (3 scales)	-.02, -.01, .15
Rolfe (1945)	47 elementary, 1- & 2-room	Citizenship course	Residual gain	3 rating scales	.36, .39, .43
Riesch (1949)	22 elementary	Achievement in social studies	Residual gain[c]	Superintendent rating	.22
		Personality	Residual gain[c]	Superintendent rating	.20
		Right conduct	Residual gain[c]	Superintendent rating	.35
		Social adjustment	Residual gain[c]	Superintendent rating	.24
		Attitude	Residual gain[c]	Superintendent rating	.81
		Composite all 5 measures	Residual gain[c]	Superintendent rating	.13

---

[a] Exclusive of 1- and 2-room schools

[b] Level of confidence of difference in mean ratings between instructors whose classes obtained grades in chemistry higher than predicted and those whose classes obtained grades lower than predicted.

[c] Tests used: Townsend & Willis Cooperative Social Studies Test
Washburne Social Adjustment Inventory
Wood Right Conduct Test
Remmers' Scale for Measuring Attitude Toward Teacher

The great discrepancies in the findings of investigators who have examined the student gains criterion emphasize the extreme variability in relationship among criteria used to indicate instructor ability. Apparently, at least within the limits of the measures so far used, the relationship between administrative opinion of a teacher's competence and the amount of subject matter that teacher will impart to her students cannot be predicted. While there may be no single measure that correlates consistently with measures of student change, it appears, as Jayne (166) has pointed out, that a composite index may be found which has high correlation with the student gains criterion.

THE PREDICTORS--TRAITS AND QUALITIES ASSUMED TO BE RELATED TO INSTRUCTOR EFFECTIVENESS

As might be expected many research investigations have been concerned with measuring or assaying those abilities, traits, qualities, and personality characteristics which are assumed to contribute to success in teaching. Assumptions are usually implicit also that the effect of a trait tends to be constant, that potential instructors can be selected on the basis of these traits, and that effective and ineffective instructors can be differentiated in terms of patterns of traits. Traits related to failure have also been investigated and are summarized in a later section.

Among the traits and qualities of teachers that have been investigated, studies most frequently have been concerned with the following characteristics: intelligence, scholastic achievement (academic level reached or grades obtained), knowledge of subject matter, age and experience, cultural

POOR ORIGINAL COPY - BEST
AVAILABLE AT TIME FILMED

background, teaching ability, teaching aptitude, professional attitude toward and interest in teaching, emotional stability and social adjustment, and personality. Attempts have been made to evaluate and relate a teacher's personality in general to teaching success and also to indicate the relationship to teaching ability of such allegedly specific personality traits as aggressiveness and control, appearance, considerateness, cooperativeness, enthusiasm, motivation, objectivity, and reliability. Some factor analysis studies have also been made (12, 70, 142, 149, 189, 220, 277, 285, 288, 289, 295, 314) in order to determine to what extent various factors contribute to teaching effectiveness.

In the following pages the available quantitative studies relative to these traits and qualities will be summarized. In considering the various correlation coefficients reported it should be remembered that their meaningfulness may be limited by the use of unvalidated criteria such as ratings, and their magnitudes may be limited by unreliabilities of the criterion as well as of the predictors.

POOR ORIGINAL COPY - BEST
AVAILABLE AT TIME FILMED

Intelligence as Related to Instructor Effectiveness

It would appear at first glance that of the desirable teacher characteristics one of the most important should be intellectual brightness. That there might be a relationship between teaching ability and intelligence was realized even before the Stanford revision of the Binet-Simon Intelligence Test popularized the I.Q. and the Army Alpha provided an easily accessible measure. This implicit hypothesis that teaching effectiveness and intelligence are related is reflected in the correlations between ratings of these two teacher variables; such correlations may be high because of halo effect, or more accurately, because of the logical error of assuming that intelligence and teacher merit are related.

In 1912 for instance, Boyce (48), basing his findings on the rankings of 328 secondary school teachers by 27 administrators, reported a correlation coefficient of .71 between ranking on general merit and ranked estimate of intellectual capacity. As late as 1929 Baird and Bates (14) secured subjective ratings of intelligence of 444 elementary school teachers made by their principals with a five-point scale. When general merit ratings were correlated with estimates of general intelligence a correlation coefficient of .58 was obtained. The corresponding coefficient for social intelligence was .57. When these coefficients are compared with those obtained by using more objective measures of intelligence (see Table 17), the presence of the halo effect in these estimates of intelligence becomes apparent.

Intelligence Test Scores as Related to Instructor Effectiveness

In 55 of the available studies which have appeared in the last 25 years, attempts have been made to relate objective measures of intelligence of the teacher to various measures or estimates of teaching effectiveness. Intelligence test scores have been correlated with practice teaching ratings or grades, various administrative ratings, student ratings, and pupil gains. In the studies mentioned, 17 different intelligence examinations (in some cases two or more) were employed. The American Council on Education Psychological Examination was used in 12 studies and the Army Alpha in 7 studies.

In 15 studies (12, 20, 52, 58, 87, 119, 125, 172, 203, 208, 247, 261, 275, 280, 323) negative correlations were reported, the largest being those of Riesch (275) $\underline{r} = -.34$, Jones (172) $r = -.26$ and Stephens and Lichtenstein (323) $\underline{r} = -.24$. All three of these coefficients were obtained when intelligence of the teacher was correlated with student gains. In 16 investigations (20, 32, 39, 56, 57, 79, 119, 133, 171, 172, 184, 185, 188, 256, 282, 320) positive correlations with $\underline{r} = .30$ or more are reported between teachers' intelligence test scores and various criteria of teacher effectiveness. The highest relationship, a correlation coefficient of .57 with student gains, was reported by Rostker (282) for a group of 28 teachers. (LaDuke in Reference 188 mentioned a coefficient of .61 in the conclusion of his study, but no zero-order coefficient of this magnitude appears elsewhere in his report. Between a composite measure of student gains and teacher intelligence he found a coefficient of .43.) Among the 55 available studies in which correlations are reported between intelligence scores and various criteria of teacher effectiveness, the number of subjects is often so small--in one instance, in part of Jones' (172) study, as few as six--that the correlation coefficients reported have little meaning.

In Table 16 are shown correlation coefficients obtained between scores on the American Council of Education Psychological Examination and several criteria of instructor effectiveness. It will be observed that the correlation coefficients reported vary from -.26 to .57. This would appear to indicate that whether or not intelligence is an important variable in the success of the teacher depends upon the situation.

In Table 17 appear the 24 studies (8 have 2 entries) in which findings are given for 90 or more teachers. The first 18 entries are concerned with student-teacher groups. With the exception of the Pyle (261), Breckinridge (52), and Fuller (125) investigations most of the studies report a low positive correlation between intelligence and practice teaching grade or rating. The last 14 entries relate to groups of teachers in the regular school situation. Except for Somers (320), Kriner (185), and Gould (133), these latter investigations appear to show that there is only a slight relationship between the intelligence and rated success of a teacher.

It was noted earlier that student grade or achievement is sometimes negatively related to the rating of teachers by students and sometimes positively, because some teachers may be better for bright students and others for dull students. Similarly, the relationship of instructor intelligence to instructor competence may be positive, negative, or nonexistent depending upon motivation and ability of students, subject matter,

Table 16

Correlations Between A.C.E. Psychological Examination and
Various Measures of Teacher Effectiveness

Investigator	Teacher sample	Effectiveness measure	Correlation
Martin (1944)	123 with 1 yr. experience	Superintendent rating Practice teaching	.02 .14
LaDuke (1945)	31 elementary, 1 room	Pupil gain	.43
Rolfe (1945)[a]	47 elementary, 1- & 2-room	Pupil gain (social studies)	-.10
Rostker (1945)[a]	28 elementary, rural[b]	Pupil gain (social studies)	.57
Seagoe (1945)	31 student	Practice teaching rating	.12
Seagoe (1946)	25 elementary, 2 yr. experience	Supervisor ranking	.10

---

[a] Rolfe (280) and Rostker (282) also found correlations between Teachers College Psychological Examination scores and pupil achievement, the coefficients being .05 and .40, respectively. In addition to the A.C.E., Jones (172) and Lins (203) used the Henmon-Nelson Intelligence Test. Jones reports correlations of .11 and .02 between the Henmon-Nelson Intelligence Test and supervisors' ratings and pupil gains respectively. Lins obtained correlations of .15, -.25, and .04 between intelligence scores on this test and the criteria stated above.

[b] Exclusive of 1- and 2-room schools.

[c] Coefficient of contingency.

Table 16 (Cont.)

Investigator	Teacher sample	Effectiveness measure	Correlation
Jones (1946)[a]	31 high school	Superintendent rating	.10
	19 high school	Pupil gain	-.26
	6 English	Pupil gain	-.24
Lins (1946)[a]	56 high school women, 1 yr. experience	Five observers' ratings	.24
	48 high school women, 1 yr. experience	Student rating	-.12
	17 high school women, 1 yr. experience	Pupil gain	-.01
Gould (1947)	98 high school, 1 yr. experience	Administrator rating	.53[c]
Stephens & Lichtenstein (1947)	20 elementary	Pupil gain	-.24
Bach (1952)	76 high school, 1 sem. experience	Practice teaching	.002
	55 high school, 1 sem. experience	Principal rating (2 different blanks)	-.01 & -.10
		2 different superintendents' ratings same scale	-.01 & -.08

---

[a] Rolfe (280) and Rostker (282) also found correlations between Teachers College Psychological Examination scores and pupil achievement, the coefficients being .05 and .40, respectively. In addition to the A.C.E., Jones (172) and Lins (203) used the Henmon-Nelson Intelligence Test. Jones reports correlations of .11 and .02 between the Henmon-Nelson Intelligence Test and supervisors' ratings and pupil gains respectively. Lins obtained correlations of .15, -.25, and .04 between intelligence scores on this test and the criteria stated above.

[b] Exclusive of 1- and 2-room schools.

[c] Coefficient of contingency.

Table 17

Correlations Between Various Psychological Examinations and Measures of Teacher Effectiveness for Groups of Teachers of 90 or More

Investigator	Teacher sample	Psychological test	Effectiveness measure	Correlation
	Student teachers			
Whitney (1922)	780	Army Alpha & Thorndike	Rating	.16
Somers (1923)	156	Composite (4 tests)	Rating	.48
Cooper (1924)	107	Thurstone	Grade	.22[a]
Hamrin (1927)	108	Army Alpha	Rating	.20
Pyle (1928)	358 (1st yr.)	Detroit Advanced	Grade	.15
	105 (advanced)	Detroit Advanced	Grade	-.02
Shultz (1928)	108	Army Alpha	Grade	.09
		Terman	Grade	.13
Markt & Gilliland (1929)	143	Otis	Grade	.07
Broom (1929)	148	Thorndike	Grade	.30
Collins (1930)	104	Thurstone	Grade	.34
Ullman (1930)	116 (senior)	Brown University	Rating	.24
Whitney & Fraiser (1930)	100	Thurstone	Rating	.24
Breckinridge (1931)	420 (beginning)	Army Alpha	Grade	.09
	(advanced)	Army Alpha	Grade	-.04
Broom (1932)	134	Thorndike	Grade	.38
	136	Thorndike	Rating	.24
Coxe & Cornell (1933)	900 (approx.)	Terman	Grade	.06
Bent (1937)	597	Miller Analogies	Rating	.20
Major (1938)	122	Ohio State	Grade	.14
Martin (1944)	123	A.C.E. Psychological Examination	Grade	.14
Fuller (1946)	93	A.C.E. Psychological Examination	Rating	-.06
	95	Miller Analogies	Rating	.02
	In-service teachers			
Whitney (1922)	780	Thorndike & Army Alpha	Supervisor rating	.03
Somers (1923)	110 with 1 yr. experience	Composite (4 tests)	Supervisor rating	.43
Hamrin (1927)	108 with 1 yr. experience	Army Alpha	Superintendent rating	.04
Pyle (1928)	99 with 1 yr. experience	Detroit Advanced	Principal estimate	.03
	99 with 2 yr. experience	Detroit Advanced	Principal estimate	.02
McAfee (1930)	98 elementary	Unknown	Supervisor rating	-.08
	112 elementary	Unknown	Observer rating	.20
Wagenhorst (1930)	191 with 1 yr. experience	Otis	Superintendent	.00
Ullman (1930)	116 high school, 1 sem. experience	Brown University	Administrator rating	.15
Breckinridge (1931)	215	Army Alpha	Administrator rating	.07
Broom (1932)	137	Thorndike	Administrator rating	.17
Coxe & Cornell (1933)	500 (approx.) elementary, 1 yr. experience	Terman	Supervisor rating	-.06
	400 (approx.) elementary, 2 yr. experience	Terman	Supervisor rating	.01
	112 elementary, 2 yr. experience	Terman	Expert observer rating	.03
Phillips (1935)	173 elementary & jr. high school	Ohio State	Superintendent & principal rating	.33
			Sigma score of rating	.26
Odenweller (1936)	560 elementary	Various	Administrator ranking	.00 & .06
Kriner (1937)	94 with 1 yr. experience	Thurstone	Superintendent rating	.35
Martin (1944)	123 with 1 yr. experience	A.C.E. Psychological Examination	Administrator rating	.02
Gould (1947)	98 with 1 yr. experience	A.C.E. Psychological Examination	Administrator rating	.53[a]

---

[a] Coefficient of contingency.

classroom conditions, and other factors. In correlating instructor intelligence with effectiveness, the assumption is implicit that the effect of intelligence is constant regardless of time, type of student, nature of subject matter, educational objectives, classroom climate, and the like. The variety of relationships found by investigators in this area provides strong support for questioning this assumption. In some cases too much intelligence on the part of the teacher may constitute somewhat of a handicap. This is understandable when one considers the possibility that some teachers may not be able to "get down" to the level of the student. In a technical school situation this might very well be the case, especially where civilians having considerable technical or academic training are employed.

Considering the more or less restricted range into which the intelligence of a public school teacher may be expected to fall (intelligence quotients with a range of 103 to 126 and an average of 114 as reported in findings with the Army Alpha[1]); for all practical purposes this variable is of little value as a single predictor of rated teacher success, inasmuch as it would be used with a population already selected on the basis of intelligence.

Although no particular relationship is shown between intelligence of teachers in general and teaching competence, it is possible that in the case of teachers of more advanced subject matter a significant relationship might be found. The investigations of Knight (178), Jones (171), Boardman (39), Ullman (339, 340), and Jones (172) who worked with high school teachers might be expected to throw some light on the possibility. With the exception of the correlation reported by Jones (172) who obtained a coefficient of -.26 when he correlated intelligence of 19 high school teachers with student gains, correlations ranged from .10 to .45, the latter coefficient being obtained by Knight (178), apparently with less than 38 subjects. It is seen that these correlation coefficients tend to be somewhat higher and somewhat less variable than those reported for elementary teachers.

In 1927 Pyle (260) pointed out ". . . we find that intelligence as determined by various types of psychological experiments is a just-barely-perceptible factor in teaching success." The studies involving groups of teachers of 90 or more which were summarized in Table 17 have largely supported this generalization to the extent that low positive correlations have usually been reported. Of 42 product-moment correlation coefficients between some measure of intelligence of the teacher and some criterion of teaching success, 37 were positive and ranged from zero to .48 while only 5 were negative, the largest of these latter being -.08.

Intelligence test scores are probably of little value as indicators of success or failure with respect to teachers of the lower academic grades. This is probably due to the narrow range of scores involved, the teachers from whom intelligence test scores have been obtained for research purposes constituting a highly selected sample of the total population. In some teaching situations the intelligence factor, however, may make some contribution when used with measures of other instructor variables as a predictive device. If one considers the mean scores of instructors teaching very diverse subject matter (e.g., calculus vs. trade school) significant differences in intelligence between instructor groups may appear. An intelligence tests score below the minimum found for an instructor of certain subject matter might well predict lack of success in teaching, for instance in the more complex levels of such a field as mathematics.

---

[1] Army Alpha scores range from 97 to 148 with an average score of 122.

In the Air Force technical schools there is some indication that intelligence may be somewhat more important as an instructor variable. Morsh and Swanson (232) reported a correlation coefficient of .46 (significantly different from zero at the .01 level) between Army General Classification Test scores and supervisors' ratings of 38 instructors of reciprocating engine courses on the Instructor Description Form (154).

The restriction of range of intelligence which may have kept the correlation coefficients low when obtained with elementary or high school teachers may not occur in the instructor population of the Air Force where the range of intelligence may be much greater than that of civilian teachers. It may be expected, however, that intelligence will bear a differing relationship to teaching success, depending upon the complexity of the course material and the level of student aptitude and experience compared with that of the instructor. Consequently, great care must be taken in generalizing from one course to another. The correlation of instructor intelligence with the criterion of student gains might well be quite different for high level courses, such as the weather courses, in which the students are highly selected, as compared with a course such as sheet metal.

Education as Related to Instructor Effectiveness

From 1905 to 1951 some 26 studies were made of the relation of amount or kind of education of a teacher to success in the classroom. In 9 of these studies statistical relationships between some criterion of instructor efficiency and amount of education were determined. These investigations have been summarized in Table 18.

Results of these studies are difficult to interpret. In the great majority of the investigations, the range of education is given but the variability in the amount of education for the teachers studied is not indicated. As in the case of intelligence the restriction of range in the amount of education tends to lower the obtained correlation. Also the criterion used in most of these studies is highly suspect and any relationship found may primarily reflect contamination in the criterion. The two highest correlations were one of .42 found by Knight (178) and one of .41 reported by Davis and French (97). In the Knight study the education

Table 18

Relation of Education to Instructor Effectiveness

Investigator	Teacher sample	Measure of effectiveness	Measure of education	Correlation
Meriam (1905)	504 elementary	Principal ranking	Amount of training	College graduate less successful in lower grades
Ruediger & Strayer (1910)	204 elementary	Supervisor ranking	Amount of training	Normal school graduate most effective; college, next, high school only, least
Boyce (1912)	300 (approx.) high school	Superintendent & principal rating	Amount of training	Comparison slightly in favor of those having professional training.
Ritter (1918)	1738	Official rating	Amount of college training	.29
Knight (1922)	52 elementary & high school	Fellow teacher rating	Amount professional study[a]	.28 elementary
		Fellow teacher rating	Amount professional study[a]	.42 high school
		Supervisor estimate	Amount professional study[a]	.38 elementary
		Supervisor estimate	Amount professional study[a]	.36 high school
Lang (1924)	154 elementary 113 high school	Supervisor rating Supervisor rating	Amount of training exclusive of professional subjects	-.11 & -.22 -.20 & -.16
	154 elementary 113 high school	Supervisor rating Supervisor rating	Amount of professional training	-.20 & -.13 -.13 & .19
Barthelmess & Boyer (1928)	5002 elementary 1220 jr. high school	Principal ranking Principal ranking	Credits[a] Credits[a]	.16 .19
Davis & French (1928)	2156	Official rating	Amount professional training	.41
Jacobs (1928)	50 good; 50 poor elementary	Principal rating	16 different college courses	Critical ratios for 8 courses favored good teacher; other 8 favored poor teacher
Moss, et al. (1929)	(Number unreported) college	Pupil gain	Ph.D. vs. M.A.	Results not significant
Broom (1932)	243	Administrator rating	Units in education	.01
Davis (1934)	1263 high school	Pupil gain	Amount of training in subject taught	50.3% of classes taught by teachers with more training excelled; 49.7% of classes did not.
Odenweller (1936)	560 elementary	Principal, assistant principal, & supervisor ranking	Courses beyond 2 yr.	.11
Sandiford, et al. (1937)	(Number unreported) male	Practice teaching grade	Graduate training	3.0 (critical ratio)
Young (1937)	1521 high school	Principal ratings	Degree held; amount of training in subject taught and education courses	Ratings tended to be higher for teachers with higher degrees and more training
Nutting (1943)	(Number unreported) elementary & high school	Principal ranking	Amount of training and degree held	No relationship
Rolfe (1945)	47 elementary, 1- & 2-room	Pupil gain	Training above high school	-.09
Nemec (1946)	265 elementary & high school	Failed to get permanent certificate at end of 2 yr.	Type of professional training	Relationship not significant at .05 level
Stephens & Lichtenstein (1947)	86 elementary	Pupil gain	Type of professional training	No difference except that those with professional training scored somewhat higher than a group of teachers with no professional training
Riesch (1949)	22 elementary	Superintendent rating Pupil gain	Years of training (1-4) Years of training (1-4)	.15 .00
Riley, et al. (1950)	372 college	Student rating	Degree held	68% Ph.D's above median 27% M.A.'s above median 25% A.B.'s above median
Morsh & Swanson (1951)	65 A.F. instructors 34 A.F. instructors	Supervisor ratings Supervisor ratings	Years of education Years of education	.08 .02
Ryans (1951)	191 elementary	Composite observers ratings	Amount of training	.11[b]

---

[a] In-service

[b] Contingency coefficient

measure was that of amount of in-service training taken and the relationship may only reflect the extent to which raters look with high favor on such training. Davis and French compared official ratings reported to a state educational department with amount of professional training. Here again the raters were probably aware of the amount of training each teacher had, and such knowledge may well have influenced their ratings.

Another source of error in studies comparing amount of education with teaching efficiency is that often the factors of age and years of teaching are not held constant. Frequently the teachers with the "poorer" educational background as defined in the different studies belong to the group of older teachers so that factors other than amount of education may be operating to result in their getting a lower rating.

Some of the studies reported are too old to have much significance for present day education. The variables, elementary teaching and college education for instance, have changed radically since 1905. The studies are of some historical interest, however, and may also be used to see if any changes have occurred. It is interesting to note that in 1905 Meriam (225) said, "Professional work in Normal Schools does not contribute as much as one would expect, though Normal School graduates do better than teachers in city training schools, and these in turn better than teachers with no professional education." Then in 1938 Allen (2) in a study of 60 superior and 60 inferior teachers makes the following similar statement, "After a relatively high minimal background has been reached in such items as are normally stressed in substantial teacher-training programs, further addition to these backgrounds are not necessarily the things which differentiate superior from inferior teachers."

In 1944 Daniel (91) reported a study in which educational levels of teachers rated "excellent" were compared with the percentage of all teachers of their state having the same educational level. He asked a large sampling of superintendents, supervisors, principals, teachers, pupils, and patrons of schools in South Carolina to indicate their "best" teachers. In Table 19 is shown the percentage of "best" teachers as indicated by pupils and patrons (parents) for the various educational levels and percentages of the teacher population for the state as a whole. Unfortunately, these data do not necessarily show that teachers with better education are really better teachers. They may have been rated "best" because of their education.

In 1951 Ryans (286) found no significant differences when 275 elementary teachers were divided into groups based on amount of college training. The criterion of teaching effectiveness was factor scores obtained when composite observer rating was factor analyzed by the centroid method. The contingency coefficient based on 191 cases was .11.

Considered as a group the investigations of semester hours or years of education as related to instructor efficiency have shown that any relationship that may exist is slight. Results of these studies suggest that further

Table 19

Educational Qualifications of "Best," White, High School Teachers[a]

Educational Level	"Best" teachers N	"Best" teachers %	South Carolina teachers %
High school graduation or less	1	0.5	0.5
2 years of college	2	0.9	0.3
3 years of college	3	1.5	0.9
Bachelor's degree	45	21.8	61.5
Bachelor's degree plus	98	47.6	24.4
Master's degree	20	9.7	10.2
Master's degree plus	37	18.0	2.0

---

[a] Taken from a study by Daniel (91).

search along lines followed here for factors which differentiate the effective from the ineffective teacher will probably not be too rewarding.

Such variables as "years of education" or "semester hours" lack meaning unless psychological or educational changes induced in individuals undergoing training can be measured. Whether or not a teacher has had a course in educational psychology has little significance because of the variation in such courses from college to college and even from instructor to instructor within a given college. We learn from these studies, what we might have suspected from the beginning, that the amounts of education or semester hours are meaningless variables in relation to measures of teacher effectiveness. Progress in research in this area can be made only when more specific and detailed measures of the effects of training are developed as variables and substituted for the gross indications of educational achievement used heretofore. More meaningful variables might be provided, for example, by using direct measures of the outcomes to be expected from given amounts of training of a specific kind such as might be associated with child psychology, psychology of learning, or other subject matter courses.

On the basis of what has been reported to date, however, it can only be said that beyond certain more or less obvious knowledge requirements, greater or lesser education of the teacher in terms of courses or semester hours seems to be unimportant. Where any substantial relationship has been shown, the possibility of contamination of data has not been eliminated since a school administrator's rating of a teacher may be influenced by what he knows about that teacher's training. There is some suggestion from the text of a number of articles that the primary motivation for research lay in the educator's enthusiasm for some particular course or combination of courses in his institution. It is thus perhaps inevitable that some of the results received somewhat less critical interpretation here than they deserved.

Scholarship as Related to Instructor Effectiveness

In the search for variables which might be used as bases for the prediction of teaching effectiveness, one of the most obvious indicators in terms of accessibility and objectivity would appear to be that of previous scholarship. The hypothesis is rather widely held that the individual who is himself a good student of mathematics, for instance, can impart his mathematical information to others. In line with this assumption, in Air Force technical schools instructors are frequently selected on the basis of grades they obtained in particular subject matter courses. Another school of thought maintains that knowledge of subject matter is not as important as knowledge of teaching methodology, thus assuming that the student teacher who excels in practice teaching or in courses in methods will automatically become a good teacher.

In the attempt to relate scholarship to teaching competence two types of studies have been made. The first of these involves the investigation of academic grades received by student teachers as they are related to standing in practice teaching. The second type concerns the competence of teachers in the school situation as related to their earlier scholarship in terms of grades received in school or college, including general scholarship, standing in academic major, professional education and methods courses, with particular emphasis on grades in practice teaching.

The usual measure of scholarship is expressed in terms of grade-point average or grade-point ratio, which is grade weighted by the number of hours or units credit in the course. In Tables 20 and 22, various designations used by investigators (general scholarship, marks, average grades, honor point ratio, academic average, etc.) have all been interpreted by the reviewers as the college scholarship variable.

Practice Teaching Grades versus Scholarship

Many attempts have been made to relate practice teaching grades to scholarship in an effort to obtain some basis for forecasting success in practice teaching. By implication, a good standing in practice teaching would indicate probable success later in the school situation itself.

Of some 31 studies of teachers in training available to the reviewers, 23 report correlations obtained between some measure of average college grades and grades or ratings in practice teaching, 16 report correlations between standing in specific college courses and practice teaching, and

9 report correlations found between high school scholarship and practice teaching. The results of these studies are summarized in Table 20. It will be noted that the correlation coefficients shown are all positive, and in several instances where comparatively large groups are involved they are quite substantial. There is greater variability in the case of the coefficients found when grades in specific courses are compared with practice teaching than when the average for all college courses is so compared. This variability probably has little meaning due to differences in sizes of groups used and in methods of obtaining the original data.

The implication is quite clear, however, that grades a student will obtain in a practice teaching course may to some extent be predicted by the grades that student obtained in college. Unfortunately, there is no indication in the studies reviewed that steps were taken to keep the measures of practice teaching experimentally independent and uncontaminated. In other words, persons assigning practice teaching grades were apparently not kept unaware of the grades obtained by the students in other college courses. This means that the positive correlations in Table 20 may be attributable in part to the operation of logical error or halo effect. The instructor who grades his student on practice teaching may give higher grades to the student he knows to have received higher grades in his previous college work. On the other hand, in the light of the positive coefficients found regardless of the course or courses correlated with practice teaching general scholarship may be the determining factor. It is probable, too, that both performance in practice teaching and general scholarship are related to intelligence level. The importance of this relationship depends, however, on the extent to which practice teaching grades predict later success as a teacher. The research on this question is reviewed in the next section.

With the exception of one study, Somers' (320), the coefficients reported for high school standing, though positive, are rather low. From this it would appear that while some positive relationship is found for groups, little prediction of success in practice teaching may be made on the basis of an individual's scholastic record in high school. Although again the investigators do not state whether or not the persons assigning the practice teaching grades were kept unaware of the student's high school standing, the probabilities are that halo effect was not present to any great extent here. It is doubtful if, in most college situations, college instructors are aware of their students' high school grades. However, it is also true that there is very little variability in the high school grades of college students, since the better students tend to go on to college. This latter factor would operate to lower the correlation coefficients obtained.

Scholarship versus Teaching Success in the Field

The second broad approach in relating scholarship to teaching ability is that of cosidering high school or college records of teachers who are

Table 20

Relation of Practice Teaching Grades or Ratings to Scholarship

Investigator	Number of student teachers	High school grades or rank	College grade average	Major subject	Educational method	Other courses
Mean & Holley (1916)	40		.24	.19	.57	
Fordyce (1919)	123		.61	.70		
Whitney (1922)	780	.27	.39		.21	
Somers (1923)	156	.64				
Cooper (1924)	107		.33[a]			
Hamrin (1927)	108		.45			
Shults (1928)	108	.08	.43			
Zant (1928)	200				.32	.30 (Psychology)
Broom (1929)	148				.21	
Morris (1929)	60		.55			
Ullman (1930)	116		.26	.22	.46	
Whitney & Frasier (1930)	100 (selected)		.47			
	70 (control)		.52			
Breckinridge (1931)	420 (beginning course)	.26				
	(advanced course)	.12				
Neal & Mead (1931)	64		.37	.49		
Broom (1932)	235 (grade)		.58			
	232 (grade)				.45	
	235 (rating)		.39		.04	
Broom & Ault (1932)	55 (rating public school)				.11	
	48 (rating public school)		.44			
	63 (rating college)		.53			
	68 (rating college)				.22	
Coxe & Cornell (1933)	900 (approx.)	.09				
Dodd (1933)	90		.35			
Hatcher (1934)	20		.25			
Butler (1935)	242		.40		.23	
	118		.46		.43	
Kriner (1935)	55	.33	.52			
Bent (1937)	597	.21	.46	.45	.27	.29 (English)
Lawton (1939)	705 (1932)		.48[b]			
	528 (1936)		.45[b]			
	497 (1937)		.46[b]			
Martin (1944)	123	.07	.12			
Hult (1945)	100		.49	.45	.51	.36 (Minor subject)
	76		.35	.14		
	67					.16 (Minor subject)
Seagoe (1945)	25		.52		.47	
	23			.51		
Fuller (1946)	85	.03				
	53			.62		
Schwartz (1950)	34		.32		.56	
Bach (1952)	76		.67		.19	

---

[a] Coefficient of mean square contingency

[b] College leaving examination

now on the job. Some 49 such research studies have been examined. The results of these investigations are discussed in the following pages under two headings: (1) Practice Teaching Grades versus Teaching Success in the Field and (2) Other Academic Grades versus Teaching Success in the Field. The latter section includes general college average; grades in major subject, education courses and other specific college courses; and high school grades or rank.

Practice Teaching Grades versus Teaching Success in the Field

In 31 available studies practice teaching grades or ratings were compared with some criterion of on-the-job teaching success. In 29 of these summarized in Table 21, the correlation coefficients ranged from -.17 to .84. The .84 coefficient was obtained by Tudhope (336) in a study of 50 male teachers in England. This investigator's data probably reflect contamination due to the rating of teachers in service by the same official inspectors who participated in assigning practice teaching grades.

As indicated in Table 21 with two exceptions, Broom and Ault (58) and Jones (172), all of the available studies reported a positive relationship between practice teaching grades and criteria of success in the field. Most of the correlation coefficients are low, however, only six being .40 or better.

Upon examination of Table 21, it will be noted that many of the investigators used a teacher population of under two years' experience. It might be expected that if grade in practice teaching was predictive of later success in teaching, a larger correlation would be found in those studies with the less experienced teachers. Presumably after about two years of experience, a selective factor has entered the picture, the failures and teachers who have not adjusted to the teaching situation having been eliminated. This hypothesis does not stand up under the results as presented in Table 21, however, as many of the studies with inexperienced teachers report extremely low correlations. In fact, those correlations reported in studies whose population included the more experienced teachers are equally as high as many reported in studies with inexperienced teachers. These results might be partially explained by the inadequacy of the criteria used. In the great majority of these studies some form of administrative rating was employed. Since there appears to be a definite tendence of administrators to withhold high ratings from beginning teachers their ratings may be forced toward the lower end of the scale, thus curtailing the range of the sample studied. In only one of the studies, Seagoe (298), were the teachers ranked rather than rated. Seagoe obtained a correlation coefficient of .49 using the criterion of teachers ranked within their own faculty, the ranks being converted to percentile scores for analysis. In two of the studies of inexperienced teachers less fallible criteria were used. Coxe and Cornell (87) reported a correlation coefficient of .28 ($\underline{N}$ = 112) for trained-observer rating while Lins (203) obtained a coefficient of .25 ($\underline{N}$ = 58) for observer rating and a coefficient of .21 ($\underline{N}$ = 17) for pupil gain when these measures were correlated with grades in practice teaching.

Table 21

Relation of Practice Teaching Grades or Ratings to Teaching Effectiveness in the Field

Investigator	Teacher sample	Measure of effectiveness	Correlation
Merian (1905)	1165 elementary	Normal school principal estimation	.44
Moody (1918)	107 men 527 women	Salary Salary	.25 .25
Whitney (1922)	780 with 1 sem. experience	Supervisor rating	.24
Somers (1923)	110 with 1 yr. experience	Principal rating	.70
Hamrin (1927)	108 with 1 yr. experience	Supervisor ratings	.06 (1st critic teacher) .23 (2nd critic teacher)
Armentrout (1928)	200 with 1 yr. experience	Superintendent rating Superintendent rating	.29 .40[a]
Pyle (1928)	99 with 2 yr. experience	Administrator rating	.15
Shultz (1928)	58 with 2 yr. experience	Superintendent rating	.12
Wagenhorst (1930)	191 with 1 yr. experience	Superintendent rating	.23
McAfee (1930)	98 elementary 112 elementary	Supervisor rating Classroom observer	.16 .26
Ullman (1930)	116 high school, 1 sem. experience	Average principal & superintendent rating	.36
Bossing (1931)	100 high school	Administrator rating	.69
Broom (1932)	238	Administrator rating	.26
Broom & Ault (1932)	38 to 63 with 1 yr. experience 29 to 38 with 1 yr. experience	Ratings sent Department Education Ratings sent College Placement	.02 to .30 -.17 to .10
Coxe & Cornell (1933)	500 (approx.) elementary, 1 yr. experience 400 (approx.) elementary, 2 yr. experience 112 elementary, 2 yr. experience	Administrator rating Administrator rating Composite observer rating	.13 .21 .28
Kriner (1935)	55 with 1 yr. experience	Administrator rating	.39
Hardesty (1935)	231	Superintendent rating	.07
Odenweller (1936)	560 elementary	Supervisor rating	.19
Kriner (1937)	42 (4-yr. course) 1 yr. experience 94 (2-yr. course) 1 yr. experience	Administrator rating Administrator rating	.40 .34
Saniford, *et al*. (1937)	242	Composite 7 inspectors	.35
Stewart (1940)	Rural (number not reported)	Superintendent rating	.21
Tudhope (1942)	93 with 3 yr. experience plus	Inspector rating	.81
Martin (1944)	123 with 1 yr. experience	Superintendent rating	.18
Seagoe (1946)	25 elementary, 2 yr. experience	Supervisor ranking (percentile)	.49
Jones (1946)	52 high school 32 high school	Supervisor rating Pupil gain	-.04 .13
Lins (1946)	58 high school women, 1 yr. experience 50 high school women, 1 yr. experience 17 high school women, 1 hr. experience	Composite rating (5 observer) Student rating Pupil gain	.25 .06 .21
Gould (1947)	113 with 1 yr. experience	Principal rating	.66[a]
Stephens & Lichtenstein (1947)	86 elementary	Pupil gain	.01
Schwartz (1950)	18 with 2 yr. experience	Supervisor rating	.06
Bach (1952)	73 high school, 1 sem. experience	Principal rating (2 different scales) Superintendent rating (2 different raters, same scale)	.06 and .20 .18 and .12

---

[a] Coefficient of mean square contingency.

As part of a study concerned with the relation of practice teaching success to other measures of teaching ability, Bach (12) in 1952, sought an answer to the question, "Is there any agreement in the factor patterns of critic teacher and principal ratings?" A device consisting of 13 items arranged on a five-point scale was used. Ratings were made by the critic teacher while the student was engaged in practice teaching. After four months in an actual teaching situation, ratings were again made by the beginning teacher's principal. As a result of factor analysis four factors were found for each of these ratings as follows: For practice teaching rating--pupil response, technical competence, relations with others, and personal appeal; for beginning teacher rating--technical competence, cooperative attitude, initiative, and personal appeal. In conclusion Bach states:

> "There is considerable agreement between two of the four common factors found in the analyses of the practice teaching and beginning teacher ratings, but there are nonetheless important differences. These two factors are interpreted as Technical Competence and Personal Appeal. The correlations between these two factors were .27 for the practice teaching rating and -.02 for the beginning teacher rating. High positive relationships are also found between three pairs of factors in the practice teaching analysis but only one large positive and three small negative relationships are found between the factors in the beginning teacher analysis. The above differences lead to the conclusion that in spite of the similarity of name in the two factors common to each analysis, critic teachers and principals are emphasizing different characteristics or abilities in the people they train and hire, or else they place different values upon and seek different combinations of the same abilities" (12).

From the results reported in this section, one could anticipate that research with Air Force personnel might show some relationships between standing in instructor training courses and subsequent performance as an instructor. If such correlations were shown for Air Force technical training school instructors, however, the information would become available too late to have much practical predictive application for the instructor sample used but might have implications for future instructor samples.

Other Academic Grades versus Teaching Success in the Field

In 35 available studies correlations are reported which are based on scholarship or grades received by teachers while students as compared with various criteria of the effectiveness of teachers in service. These investigations have been summarized in Table 22.

With respect to general college average the correlation coefficients, with the exception of 4 studies, Meriam (225), Coxe and Cornell (87), Jones (172) and Bach (12), are all positive but range from zero (Broom and Ault in Reference 58) to .73 (Somers, Reference 320). For the most

Table 22

Relation of Scholarship to Teaching Effectiveness in the Field

Investigator	Teacher sample	Measure of effectiveness	Scholarship: High school	Scholarship: College grade average	Scholarship: Major	Scholarship: Education courses	Scholarship: Other courses
Meriam (1905)	1185 elementary	Normal school principal estimation					.42 (Psychol.)
	(Number unreported) elementary			-.09			
Moody (1918)	527 women	Salary		.25			
	107 men	Salary		.35			
Ritter (1918)	1436 elementary & high school	Official rating		.65			
Whitney (1922)	780 with 1 sem. experience	Administrator rating	.09	.07			
Knight (1922)	19 elementary	Peer rating		.15			
	6 high school	Peer rating		.60			
	53 elementary & high school	Peer rating		.33			
Jones (1923)	45 high school women	Supervisor rating		.46[a]			
	45 high school women	Supervisor rating		.45[a]			
	44 high school men	Supervisor rating		.29[a]			
Somers (1923)	110 with 1 yr. experience	Principal rating	.77	.73			
Hamrin (1927)	108 with 1 yr. experience	Supervisor rating		.05			
McAfee (1930)	98 elementary	Superintendent rating		.15			
	112 elementary	Observer rating		.40			
Ullman (1930)	116 high school, 1 sem. experience	Administrator rating		.30	.20	.30	
Wagenhorst (1930)	191 with 1 yr. experience	Administrator rating		.01			
Anderson (1931)	480 teaching certificate	Superintendent rating	.10	.19			
	110 Bachelor Degree	Superintendent rating	.22	.21			
Bossing (1931)	100 high school	Administrator rating		.17		.19	
Breckinridge (1931)	215	Principal rating	.35				
Kriner (1931)	262 elementary & high school	Supervisor rating	.39[b]				
	184 elementary & high school	Supervisor rating	.62[b]				
	38 elementary & high school	Supervisor rating	.81[b]				
Broom (1932)	240	Administrator rating		.19			
	237	Administrator rating			.19		
Broom & Ault (1932)	81 with 1 yr. experience	Administrator rating (official)		.13		.24	
	50 with 1 yr. experience	Administrator rating					
	46 1 yr. experience	for college placement		.00		-.06	
Coxe & Cornell (1933)	500 (approx.) elementary 1 yr. experience	Supervisor rating		-.03[a]			
	400 (approx.) elementary 1 yr. experience	Supervisor rating		-.01[a]			
	112 elementary 2 yr. experience	Composite observer rating	.08	.10[a]			
Peterson, *et al*. (1934)	63	Supervisor rating		.12			
	47 to 104	Salary		.22 to .71			
Kriner (1935)	55 with 1 yr. experience	Supervisor rating	.36	.49			
Phillips (1935)	173 elementary & jr. high school	Average administrator rating		.19			
Hardesty (1934)	231	Superintendent rating		.15		.09	
Odenweller (1936)	560 elementary	Administrator	.08	.29		.26	

---

[a] Jones (1923) used senior grades and Coxe and Cornell (1933) used second semester achievement as measures of scholarship.

[b] Based on grades ($r = .39$); based on students placed scholastically in approximate top half of class ($r = .62$); based on students definitely placed in top half of class ($r = .81$).

[c] Coefficient of mean square contingency.

Table 22 (Cont.)

Investigator	Teacher sample	Measure of effectiveness	Scholarship				
			school	College grade average	Major	Education courses	Other courses
Kriner (1937)	42 in 4-yr. course, with 1 yr. experience	Supervisor rating	.27	.45		.40	.48 (Science) .23 (English) .13 (Social Studies)
	94 in 2-yr. course, with 1 yr. experience	Supervisor rating	.33	.40		.33	.47 (Science) .28 (English) .22 (Social Studies)
Sandiford, et al. (1937)	242	Inspector rating		.25		.19	.20 (English) .13 (History) .20 (Geography)
	84						.24 (Specialists)
Stewart (1940)	193 rural	Superintendent rating		.22			
	71 rural	Superintendent rating	.33				
Martin (1944)	123 with 1 yr. experience	Superintendent rating	.07	.15			
Jones (1946)	54 high school	Principal rating			.05		
	51 high school	Principal rating		.24			
	50 high school	Principal rating				.40	
	43 high school	Principal rating	.13				
	33 high school	Pupil gain			-.08		
	32 high school	Pupil gain				.26	
	30 high school	Pupil gain		-.08			
	28 high school	Pupil gain	-.22				
	10 English	Pupil gain	-.43				
Lins (1946)	58 high school women, 1 yr. experience	Composite administrator		.31	.23	.29	.35 (minor subject)
	55 high school women, 1 yr. experience	Composite administrator	.33				
	50 high school women, 1 yr. experience	Pupil evaluation		.03	.05	.13	.01 (minor subject)
	48 high school women, 1 yr. experience	Pupil evaluation	.06				
	17 high school women, 1 yr. experience	Pupil gain		.53			
	16 high school women, 1 yr. experience	Pupil gain	.69		.55	.52	.44 (minor subject)
Seagoe (1946)	25 elementary, 2 yr. experience	Supervisor rank (percentile)		.03	-.15	.01	
Gould (1947)	113 with 1 yr. experience	Principal rating		.44[c]			
Stephens & Lichtenstein (1947)	86 elementary	Pupil gain		.01			-.13 (Introduction to teaching) .01 (Education psychology) .19 (History of education) .15 (Methods of teaching reading)
Esp[illegible]schade (1948)	46 physical education, 1 yr. experience	Principal rating		.12	.24		
Schwartz (1950)	18 with 2 yr. experience	Supervisor rating		.24		.02	
Bach (1952)	70 high school, 1 sem. experience	Principal rating (2 different scales)		-.01		.09	
				-.06		-.02	
		Superintendent rating (2 different raters, same scale)		.08		-.08	
				.00		-.01	

---

[a] Jones (1923) used senior grades and Coxe and Cornell (1933) used second semester achievement as measures of scholarship.

[b] Based on grades ($r = .39$); based on students placed scholastically in approximate top half of class ($r = .62$); based on students definitely placed in top half of class ($r = .81$).

[c] Coefficient of mean square contingency.

part the coefficients tend to be low. In only 9 studies were they as great as .40 or above, and even within some of these studies great variation is shown in the size of coefficients obtained, e.g., Knight (178), Jones (171), McAfee (208), Peterson et al. (254), Lins (203). While the over-all results are not such as to permit any very confident interpretations, it would appear that some relationship exists. It may be suspected that the common relationship of general intelligence to both academic and teaching success is involved.

In two studies critical ratios rather than correlation coefficients were reported. In 1937 Stuit (326) found average college grades of 100 "superior" teachers as rated by superintendents and principals to be significantly higher than for 46 "poor" teachers (CR 2.8). Shannon (305) who, in 1940, compared 111 "highly successful," 111 "average," and 37 "failing" teachers selected from among teachers who were graduated from a state teachers college during the period 1898 to 1934, also found success in the field to be related to college scholarship (CR's 2.3 to 8.2).

In 16 of the studies reported in Table 22, investigators attempted to determine whether or not teaching effectiveness in the field might be predicted from achievement in one or more college courses apart from practice teaching. Correlations between field performance of a teacher and his grades in specific college courses yielded coefficients which tended to be low but positive. In only five investigations, Jones (172), Broom and Ault (58), Seagoe (299), Stephens and Lichtenstein (323), and Bach (12), are negative coefficients reported, these appearing among positive relationships also found in these same studies. The results in the case of specific courses appear to be much the same as those obtained when practice teaching grade or rating is compared with teaching effectiveness in the field.

The relationship of high school grades or ranks to success in teaching was studied in 13 of the investigations. As will be seen from Table 22 the correlation coefficients (except for those reported in the Jones' 1946 study) are all positive but vary from .07 to .81. The relatively high coefficients reported by Somers (.77), Kriner (.81 and .62), and Lins (.69) appear to be somewhat out of line with results obtained by other investigators.

In the great majority of the studies concerned with the relationship of scholarship and teaching effectiveness, the question of whether or not ratings by administrators were influenced by knowledge of the teachers' college scholastic record is not considered. It should be pointed out that in the case of supervisors' ratings no investigator could be certain just what knowledge might contaminate the criterion nor could this be controlled. The question concerning contamination of ratings by knowledge of high school grades should also be raised but the probability is remote, however, that many supervisors are aware of the high school grades of the teachers they rate.

Considerable effort has been expended by investigators in attempting to discover the relationships existing between on-the-job performance of teachers and earlier scholarship as reflected in over-all achievement in high school or college or standing obtained in specific college courses. The outcome of all of this research appears to be that there is some relationship but that it is probably small. So far none of these investigators has shown that the attainment of a particular standing in high school or college or the mastery of any single course or group of courses is essential to teaching competence. General college scholarship and scholarship in specific college courses are both correlated to some extent with practice teaching grades. Intelligence test scores are also correlated with practice teaching grades. Investigators have apparently treated subject matter knowledge as if it were a discrete variable. Scholarship in specific college courses, however, is probably just a less reliable measure of general scholarship, or perhaps somewhat more indirectly, of intelligence. Zero-order correlations will not indicate whether subject matter knowledge per se is related to teaching effectiveness or whether subject matter knowledge, general college scholarship, and intelligence are interrelated variables. The lack of any substantial communality of content objectives of courses that bear the same title under different instructors or in different colleges makes it unlikely that a course selected by title only will be found essential to teaching competence.

Age and Experience as Related to Instructor Effectiveness

The relations of age and of experience to instructor effectiveness are reviewed together because of the obviously close relationship between these two variables. In 1928 for instance, Bathurst (23) obtained a coefficient of .88 when he correlated them.

In Table 23 are listed 17 studies in which correlation coefficients have been reported. (Bathurst's study is included since he used Knight's Professional Aptitude Test not as a measure of "aptitude" but as a criterion of teaching effectiveness.) It will be noted that these coefficients range from -.38 to .53. This suggests either that the importance of age and experience in teaching effectiveness depends upon the particular teaching situation involved or that product-moment correlations provide an inadequate indication of any nonlinear relationships that may exist.

That the relationship between age or experience and estimates of instructor effectiveness may be curvilinear is suggested by the studies of Ruediger and Strayer (283), Young (362, 363), and Davis (96). Ruediger and Strayer, in 1910, used supervisors' estimates of 204 elementary teachers while Young, as reported in 1937 and 1939, used principals' ratings of 1521 teachers. These investigators reported improvement in instructor effectiveness up to 5 years, no improvement from 5 to 20 years, and some decline thereafter. Davis, in 1934, on the basis of an investigation involving approximately 1700 high school teachers, his criterion being pupil success in passing State Board tests, concluded that pupils taught

Table 23

Age and Experience as Related to Teaching Effectiveness

Investigator	Teacher sample	Age or experience	Measure of teacher effectiveness	Correlation
Meriam (1905)	387 elementary	Experience (0 to 16 yrs.)	Supervisor ranking	.10
Knight (1922)	(Number unreported) elementary & high school	Age	Fellow teacher rating	.14
		Age	Supervisor rating	.03
		Experience	Fellow teacher rating	-.04
		Experience	Supervisor rating	.14
Somers (1923)	110 with 1 yr. experience	Age	Supervisor rating	.07
Lang (1924)	154 elementary	Experience (present school only)	Supervisor rating	.26 & .39
	113 high school	Experience (present school only)	Supervisor rating	.46 & .42
Boardman (1928)	88 high school	Age	Composite ranking (supervisor, associate teacher & pupil rating)	.34
		Experience		.39
Barthelmess & Boyer (1928)	5002 elementary	Experience (0 to 30 yrs.)	Principal ranking	.27
	1220 jr. high school	Experience (0 to 30 yrs.)	Principal ranking	.36
Davis & French (1928)	2156	Experience	Official rating	.23
Bathurst (1928)	171 high school	Age	Knight Professional Aptitude Test	.06
		Experience	Knight Professional Aptitude Test	.15
		Age (experience factored out)	Knight Professional Aptitude Test	-.15
		Experience (age factored out)	Knight Professional Aptitude Test	.21
Bathurst (1929)	300 elementary	Age	Knight Professional Aptitude Test	-.03
		Experience	Knight Professional Aptitude Test	.06
		Age (experience factored out)	Knight Professional Aptitude Test	-.17
		Experience (age factored out)	Knight Professional Aptitude Test	.18
Odenweller (1929)	560 elementary	Age (18 to 46 yr.)	Ranking (supervisor, principal, assistant principal)	.15
		Experience (1 to 7 yr.)		.15
Kriner (1931)	262 (131 best & 131 worst)	Experience	Superintendent opinion (elementary teachers)	.10[a]
		Experience	Superintendent opinion (high school teachers)	.26[a]
		Experience	Superintendent opinion (total group)	.18[a]
Rolfe (1945)	47 elementary, 1- & 2-room	Age (20 to 54 yr.)	Residual pupil gain ([illegible] and 8th grade)	.01
		Experience (1 to 30 yr.)		[illegible]
Jones (1946)	54 high school	Experience	Supervisor rating (Wisconsin M-Blank)	-.07
	33 high school	Experience	Supervisor rating (Wisconsin M-Blank)	.04
Stephens & Lichtenstein (1947)	40 (approx.) elementary, normal school grade	Age	Pupil achievement quotient	.41
		Experience (0 to 9 yr.)	Pupil achievement quotient	.53
	23 (approx.) elementary, city school grade	Age	Pupil achievement quotient	-.38
		Experience (4 to 24 yr.)	Pupil achievement quotient	-.21
Riesch (1949)	22 elementary, city & rural	Age (20 to 68 yr.)	Supervisor rating (Wisconsin M-Blank)	.14
		Experience (1 to 43 yr.)	Supervisor rating (Wisconsin M-Blank)	.35
		Age (20 to 68 yr.)	Residual pupil gain	-.01
		Experience (1 to 43 yr.)	Residual pupil gain	.00
	18 elementary, rural	Age	Supervisor rating (Wisconsin M-Blank)	.08
		Experience	Supervisor rating (Wisconsin M-Blank)	.11
Ryans (1951)	203 elementary	Experience (divided into groups of 1 to 4 yr., 5 to 9 yr., 10 or more yr.)	Composite observer rating	.21[b]

---

[a] Pearson cos π *r* coefficients.

[b] Coefficient of contingency.

by teachers with one years' experience but no better than pupils taught by teachers with two years of experience.

In 1929 Birkelo (36) using student ratings of elementary and high school teachers apparently showed increased instructor effectiveness with age, a result in agreement with that found by Daniel (91) in 1944. The significance of these findings as well as those of Ruediger and Strayer (283) and Young (362, 363) just mentioned is somewhat doubtful, however, since the proportion of each age or experience group in the total samples used is unknown.

In the few attempts to study the relationship between length of time teacher was employed in the school and efficiency ratings, higher correlations were found, as might be expected. In 1924 Lang (193) reported correlations ranging from .26 to .46 between supervisory rating of teaching efficiency and the teacher's local experience. In 1934 Davis (96) in a study of teaching efficiency based on the per cent of each teacher's pupils passing state tests in high school subjects stated that teachers with longer tenure in a given school were more successful in passing pupils through state tests than were teachers who had been employed in the same school for a shorter period of time. However, the schools which had the highest percentage of pupils passing the state tests were those schools with markedly high teacher turnover. Because of these confusing results Davis concludes, "It would seem more likely that the tenure of the teacher is a result of her success as measured by State Board tests than that success in State Board tests is a result of increased tenure." In 1945 Brookover (55) found that length of acquaintance with pupil and length of time teacher had taught in the schools, as well as age of teacher, were positively related to pupil ratings.

Several investigators in this area reported no significant differences. In 1936 Heilman and Armentrout (148) reported results of ratings on the Purdue Scale of 46 college teachers by 2115 students in 50 classes. In terms of experience teachers were divided into four groups, 7 to 12 years of experience, 12 to 17, 17 to 27, and 27 or more years of experience. Instructors were also divided into age groups by five-year intervals. No reliable differences in rating scores were found in either case. In 1946 Blair (37) compared 92 teachers with less than 10 years of experience with 113 teachers with 10 or more years of experience in terms of the number of "poor" answers on the multiple-choice Rorschach test. He also compared 107 teachers under 35 years of age with 98 teachers over 35 years of age. Differences were not significant in either comparison.

Englehart and Tucker (108), in 1936, asked 224 high school pupils to choose their best and worst teachers and to check their appropriate traits on a list. Their findings with respect to age are summarized as follows:

Age	Good teachers		Poor teachers	
	No.	%	No.	%
20 to 29	28	23.9	27	25.3
30 to 39	68	58.1	47	43.9
40 to 49	16	13.7	26	24.3
50 or above	5	4.3	7	6.5

No significance test with respect to the differences in percentages was applied. In 1946 Nemec (244) made a study of a group of 265 probationary teachers who failed to receive certificates at the end of a two-year probationary period because of unfavorable supervisory reports. When these teachers were divided into two groups (ages 19 to 22 and 23 years and over) according to the age at which they began teaching, Nemec found no differences which were significant at the .05 level. Ryans (286) in a factor analysis study of trained observers' ratings of teachers on the basis of directly observable teacher behaviors found that teachers (N = 60) with 1 to 4 years of experience were significantly different from teachers (N = 32) with 5 to 9 years of experience at the .01 level for two factors, which he named "controlled pupil activity and business-like approach" and "teacher calm and consistent, liked because human," and for the total rating. Differences were significant at the .05 level for two factors he called "pupil participation and teacher open-mindedness" and "sociability." The teachers with 5 to 9 years of experience were significantly different from the teachers (N = 111) with 10 or more years of experience at the .01 level for factors "pupil participation and teacher open-mindedness" and "teacher calm and consistent, liked because human" and at the .05 level for total rating by the observers. The teachers with 1 to 4 years of experience were significantly different from the teachers with 10 or more years of experience at the .01 level for "controlled pupil activity and business-like approach."

The research findings of Davis (96), Meriam (225), Ruediger and Strayer (283), Ryans (286), and Young (362, 363) imply that teaching effectiveness bears a curvilinear relationship to age or experience. The zero or near zero correlation coefficients reported by Bathurst (23, 24), Jones (172), Knight (178), Odenweller (247), Riesch (275), Rolfe (280), instead of showing lack of relationship, probably indicate the inapplicability of the Pearson product-moment correlation method to the nonrectilinear data involved. It appears that a teacher's rated effectiveness increases at first rather rapidly with experience and then more slowly up to 5 years or beyond. There is then a leveling off and the teacher may show little change in rated performance for the next 15 or 20 years, after which, as in most occupations, there tends to be a decline. It must be borne in mind, however, that ratings in such studies as the foregoing may suffer from the "logical error" which results from an implicit assumption that the young, inexperienced teachers can not be as good as those of 5 or more years of experience.

In interpreting the alleged decline in teaching effectiveness after 20 years or more of experience, the effect on ratings of the physical and mental changes accompanying aging in general must be considered. It is quite conceivable that while the ratings of students and supervisors might favor the younger and more vivacious teacher, the real effectiveness of teachers in bringing about student changes might not be related to age at all. There are as yet, however, no adequate studies of this relationship.

The research findings on age and experience have some interesting and rather important implications for the Air Training Command. In a study of the correlates of instructor morale in Air Force technical schools, Richey and Berkshire (273) reported percentages with respect to experience of 3117 military and 797 civilian instructors as shown in Table 24. If more valid techniques eventually confirm the findings of previous investigations that an instructor continues to improve for the first five years, the great majority of military instructors have not reached the period of greatest effectiveness. The present rotation policy may be manifestly working against best utilization of instructor potentiality in Air Force technical schools in that military personnel are not permitted to function as instructors long enough for them to achieve maximum efficiency. Any interpretation of the results of these studies for the military situation, however, must take into account the fact that military instructors may repeat the same subject matter as many as 25 times a year as contrasted with public school teachers who repeat the same subject matter only once or twice a year. It thus may well be that military instructors reach their peak in a shorter period of time than public school instructors.

Table 24

Teaching Experience of Military and Civilian Instructors In Air Force Technical Schools[a]

Experience	Military	Civilian
Less than 6 mos.	24.2%	5.4%
6 mos. to 1 yr.	41.3	11.8
1 or 2 yr.	25.4	18.9
3 or 6 yr.	7.0	17.5
5 yr. or more	2.1	46.4

---

[a] From Richey and Berkshire (273).

Knowledge of Subject Matter, Present Professional Information, And Teacher Examination Scores as Related to Instructor Effectiveness

Knowledge of Subject Matter as Related to Instructor Effectiveness

It is frequently stated that the good teacher is the one "who knows his stuff," that knowledge of subject matter being taught is the prime requisite of teaching success. With respect to this hypothesis the reviewers considered the findings of some 20 studies where various criteria of instructor competence were correlated with one or more measures of professional information or subject matter knowledge.

Much variability is evident among the coefficients found when scores on subject-matter tests are correlated with criteria of instructor competence. As shown in Table 25, these vary from -.69 to .58. It would appear that whether or not knowledge of subject matter is related to instructor competence is a function of the particular teaching situation. The negative relationships found in some studies suggest that too much knowledge on the part of the teacher may result in teaching "over the heads" of the students.

Two minor studies are not included in Table 25 because correlation coefficients were not computed. Madsen (213) in 1927, found that in terms of scores received on a test of elementary grade subjects, all except 1 of 31 teacher failures were found to be in the lowest 10% of a group of teachers studied. Allen (2), in 1938, using a test that included subject-matter knowledge, reported a low relationship between test results and teacher success for a group of 60 very superior and 60 very inferior teachers as rated by three supervisors. Only language usage and spelling significantly differentiated superior from inferior teachers.

Professional Information as Related to Instructor Effectiveness

On the basis of the nine available studies which have been summarized in Table 26, scores on tests of professional information tend to bear some slight relationship to several measures of instructor competence. With two exceptions, Rolfe (280) and Stephens and Lichtenstein (323), all the coefficients are positive. However, only two investigators, Crabbs (89), Betts (32), report any coefficients greater than .40.

National Teacher Examination Scores as Related to Instructor Effectiveness

Flanagan (112), in 1941, obtained a correlation coefficient of .51 between scores on the Common Examination of the National Teacher Examination and superintendents' ratings. He also reports coefficients significant at

Table 25

Relation of Scores on Subject-Matter Tests to Measures of Instructor Effectiveness

Investigator	Teacher sample	Test	Measure of effectiveness	Correlation
Betts (1933)	61 elementary	Subject-matter vocabulary	Pupil gain (reading)	.08 to .46
Coxe & Cornell (1933)	500 (approx.) elementary, 1 yr. experience	Tressler English	Supervisor rating	.00
		Whipple Reading	Supervisor rating	-.02
	400 (approx.) elementary, 2 yr. experience	Tressler English	Supervisor rating	.06
		Whipple Reading	Supervisor rating	.04
	112 elementary, 2 yr. experience	Tressler English	Composite observer rating	.12
		Whipple Reading	Composite observer rating	.18
Barr, et al. (1935)	66 elementary	Stanford Arithmetic	Pupil gain (arithmetic) A.Q.	.12
		Stanford Arithmetic	Pupil gain (comp. A.Q. & raw gain)	-.02
		Stanford Arithmetic	Superintendent rating (comp. 7 scales)	.02
Kriner (1935)	55 with 1 yr. experience	Cross English	Supervisor rating	.50
		Coop. English	Supervisor rating	.40
		Coop. Literary Acquaintance	Supervisor rating	.20
		Coop. General Science	Supervisor rating	.51
Kriner (1937)	42 (4-yr. course) 1 yr. experience	Cross English	Supervisor rating	.31
		Coop. English, Literature, General Science	Supervisor rating	.31, .33, .30
	94 (2-yr. course) 1 yr. experience	Cross English	Supervisor rating	.35
		Coop. English, Literature, General Science	Supervisor rating	.28, .15, .28
Martin (1944)	123 with 1 yr. experience	Coop.: Literature, Fine Arts, Science, Social Studies, Mathematics	Superintendent rating	.10, .17, .03 .09, .15
		Teacher College English	Superintendent rating	.07
Rolfe (1945)	47 elementary, 1- & 2-room	American Country Civics & Government	Pupil gain (citizenship)	-.01
		Hartmann-Info. Pub. Probs.	Pupil gain (citizenship)	.01
Rostker (1945)	28 elementary, rural	American Country Civics & Government	Pupil gain (social studies)	.36
		Wrightstone (Research Ability)	Pupil gain (social studies)	.58
Jones (1946)	48 high school	Reading Comprehension	Supervisor rating	.12
	31 high school	Reading Comprehension	Pupil gain (various subjects)	.13
	13 high school (English)	Reading Comprehension	Pupil gain (English)	-.69
Lins (1946)	44 high school women, 1 yr. experience	Coop. English	Composite supervisor rating	.22
	37 high school women, 1 yr. experience	Coop. English	Pupil evaluation	-.32
	11 high school women, 1 yr. experience	Coop. English	Pupil gain (various subjects)	.01
	57 high school women, 1 yr. experience	Coop. Reading	Composite supervisor rating	.24
	49 high school women, 1 yr. experience	Coop. Reading	Pupil evaluation	-.34
	17 high school women, 1 yr. experience	Coop. Reading	Pupil gain (various subjects)	.10
Seagoe (1946)	25 with 2 yr. experience	Coop. Mathematics, Natural Science, Social Studies, Contemporary Affairs	Supervisor ranking	-.08, -.24, .13, .12
Gould (1947)	113 with 1 yr. experience	Coop. Contemporary Affairs	Principal rating	.38[b]
Stephens & Lichtenstein (1947)	20 elementary	Arithmetic Fundamentals, Arithmetic Reasoning	Pupil gain (arithmetic)	-.52, -.04
		Reading Comprehension; English Usage, Sentence Structure	Pupil gain	-.41, -.10, -.20
		Spelling	Pupil gain	-.15

---

[a] Exclusive of 1- and 2-room schools.

[b] Coefficient of mean square contingency.

Table 26

Relation of Scores on Professional Information Tests to Measures of Instructor Effectiveness

Investigator	Teacher sample	Test	Measure of effectiveness	Correlatio
Crabbs (1925)	(Number unreported) elementary	Steele-Herring	Pupil gain	.05
			Supervisor ranking	.41
Boardman (1928)	88 high school	Professional Information (unpublished)	Composite rank (supervisor, teacher, pupil)	.26
		Procedures		.28
Ullman (1930)	116 high school, 1 sem. experience	Odell (principles of teaching)	Average superintendent & principal rating	.14
		Weber (objectives of teaching)	Average superintendent & principal rating	.09
Betts (1933)	61 elementary	Professional Information (composite 16 tests)	Pupil gain	.10 to .46
Barr, et al. (1935)	66 elementary	Torgerson Professional Information	Pupil gain A.Q.	.23
		Torgerson Professional Information	Pupil gain (composite A.Q. & raw gain).	.08
		Torgerson Professional Information	Superintendent rating (composite 7 scales)	.16
Martin (1944)	123 with 1 yr. experience	Teacher College Elementary	Superintendent rating	.02
Rolfe (1945)	47 elementary, 1- & 2-room	Lewerenz-Steinmetz (education orientation)	Pupil gain	-.06
Rostker (1945)	28 elementary, rural	Lewerenz-Steinmetz (education orientation)	Pupil gain	.30
Stephens & Lichtenstein (1947)	35-42 elementary (normal school)	Professional examination	Pupil gain	-.11
	21-26 elementary (city training school)	Professional examination	Pupil gain	-.49

---

[a] Exclusive of 1- and 2-room schools.

the .05 level between total scores on the Common Examination of the National Teacher Examination and the proportion of students reporting the particular teacher's name in response to the question: "Which teachers seemed to have a broad knowledge of other subjects besides the one you had with them?" On the other hand, when Lins (203), in 1946, correlated National Teachers Examination scores with pupil evaluation of their teachers he obtained a correlation coefficient of -.30 significant at the .01 level of confidence. When Lins used a composite gain criterion he found a coefficient of .45. The latter figure, however, is probably not significant since only seven teachers were involved.

In 1951 Ryans (286) correlated scores obtained by 192 elementary and 165 secondary teachers on the General Principles and Methods of Teaching test of the 1949 National Teachers Examination Battery with two kinds of ratings made by principals. For the elementary teachers the correlation coefficients obtained between examination scores and principals' ratings on an observation blank was .17, and between examination scores and principals' ratings of over-all effectiveness, .23. The corresponding coefficients for the secondary teachers were .13 and .15. The principals' ratings on the two blanks correlated .83 for both groups of teachers. When an analysis was made of examination scores obtained by the upper and lower 27% of the teachers, differences significant at the .01 level were obtained with respect to 52 "high" and 52 "low" elementary teachers, but the differences were not significant at the .05 level for the 45 "high" and 45 "low" secondary teachers.

Despite the more or less unpromising results that have been reported by investigators of the relationship of professional information and knowledge of subject matter to instructor effectiveness, this might still be a field in which useful research work can be done. The restriction in range of information of elementary school teachers might account for some of the low correlation coefficients. It is possible, too, that the particular subject matter involved may be a factor in determining the relationship between an instructor's competence and his knowledge of subject matter and/or professional information. It appears that in teaching certain technical school subjects, at least, the amount of technical information possessed by the instructor may be important. Morsh and Swanson (232) in a small exploratory study found a correlation coefficient of .45 (significantly different from zero at the .01 level of confidence) between power plant proficiency examination scores and supervisors' ratings of 73 instructors on a forced-choice form.

An Air Force technical school instructor must possess a certain minimum of technical information. He must be familiar with certain facts, must possess the requisite skills, and must understand the procedures involved in the specialty he is teaching in order to impart these facts, skills, and techniques to his students. The differential between instructors' knowledge as compared with that of their students is also an important consideration. The instructor with wide experience and background or technical information which goes far beyond that of his students may have the same difficulty as that of the overly intelligent instructor in communicating at the student level. On the other hand, an instructor who has the bare minimum of the knowledge requirements may be put in an embarrassing position or may actually lose the respect of older, experienced students who know more than the instructor about the subject at hand. The extent and implications of the differences between subject-matter knowledge of instructors and the knowledge of their students may vary from course to course in ways only to be determined through investigation.

Extracurricular Activities and General Culture Test Scores Versus Instructor Effectiveness

Extracurricular Activities

There is rather widespread belief among school administrators that a teacher who has taken part in activities outside the classroom in high school or college thereby becomes a more rounded person and makes a better teacher. In two investigations (292, 305) critical ratios were computed between teaching effectiveness of groups of teachers who as students had participated in extraclassroom activities as compared with teachers who had been nonparticipants.

Sandiford et al. (292), in 1937, compared the top and bottom third of the group when 336 student teachers were ranked according to teaching grades. Significant critical ratios favoring the top third were found with respect to several extracurricular activities. In terms of number of extracurricular participations, Shannon (305), in 1940, reported significant critical ratios when 86 most successful men teachers were compared with 24 failures and when 111 most successful men and women teachers were compared with 37 failures.

Since the less able student cannot keep up with his studies if he participates and hence refrains from participation or is not allowed to participate in extracurricular activities, it is necessary to partial out scholastic ability if the relationships found by Sandiford, Shannon, and others are to be attributed to the student's becoming a "more rounded person." As they stand, these results merely reflect the tendency for the brighter students both to get higher grades in all college subjects (including student teaching) and to participate more in extracurricular activities.

Several investigators (171, 182, 185, 196, 218, 298, 299, 319, 320, 324, 344) have reported correlations found between teacher or student teacher participation in extracurricular activities and ratings of teaching effectiveness. As will be seen from Table 27, in the nine studies of teachers on the job the correlation coefficients range from -.06 to .46. In general, investigators found low positive relationships between extracurricular activity and instructor effectiveness. On the basis of the results of the studies reviewed, there appears to be slight justification for further search for selection or evaluation measures in terms of the amount of extracurricular participation of a teacher while a student in high school or college.

General Culture Test Scores

Six investigators attempted to correlate scores on the Cooperative General Culture Test with measures of teacher competence.. The results are markedly inconsistent, with a rather strong negative relationship being indicated in several instances. These studies are summarized in Table 28. In addition to the studies reported in Table 28, several investigators (125, 161, 184, 218, 298) correlated total scores on the Cooperative General Culture Test with student teaching grades. Correlation coefficients obtained ranged from -.02 in the Seagoe study (298) with 31 student teachers to .21 in the Kriner study (184) with 55 student teachers. The studies reviewed appear to indicate that the relations of Cooperative General Culture Test scores to instructor effectiveness differ little from those reported for other subject matter tests.

Table 27

Relation of Extracurricular Activities to Instructor Effectiveness

Investigator	Teacher sample	Type of activity	Measure of effectiveness	Correlation coefficient
Jones (1923)	45 high school women (1920)	Extracurricular	Supervisor rating	.05
	45 high school women (1921)	Extracurricular	Supervisor rating	-.06
	44 high school men (1920-21)	Extracurricular	Supervisor rating	.27
Somers (1923)	110 with 1 yr. experience	Extracurricular	Principal rating	.41
Kriner (1931)	72 elementary & high school	Extracurricular	Supervisor rating	-.04
		Held student office	Supervisor rating	.10
Soderquist (1935)	482 adult education	277 who held student office vs. 205 who did not	Superintendent rating	.25[a]
Kriner (1937)	42 (4-yr. course) 1 yr. experience	Extracurricular	Supervisor rating	.46
	94 (2-yr. course) 1 yr. experience	Extracurricular	Supervisor rating	.25
Stewart (1940)	145 rural	Extracurricular	Superintendent rating	.24

---

[a] Bi-serial r.

Table 27 (Cont.)

Investigator	Teacher sample	Type of activity	Measure of effectiveness	Correlation coefficient
Martin (1944)	123 with 1 yr. experience	Extracurricular	Supervisor rating	.22
		Number offices held	Supervisor rating	.18
Seagoe (1946)	25 elementary, 2 yr. experience	Officer membership ratio	Administrator rating	.16
		Memberships	Administrator rating	.06
Von Haden (1946)	58 high school women, 1 yr. experience	High school extracurricular	Supervisor rating	.19
	50 high school women, 1 yr. experience	High school extracurricular	Pupil evaluation	.17
	17 high school women, 1 yr. experience	High school extracurricular	Pupil gain	.06

---

[a] Bi-serial r.

Table 28

Relation of Scores on the Cooperative General Culture Test To Measures of Instructor Effectiveness

Investigator	Teacher sample	Measure of effectiveness	Correlation coefficient
Kriner (1935)	55 with 1 yr. experience	Supervisor rating	.30
Kriner (1937)	94 (2-yr. course) 1 year experience	Supervisor rating	.25
	42 (4-yr. course) 1 year experience	Supervisor rating	.22
Martin (1944)	123 with 1 yr. experience	Superintendent rating	.11
Seagoe (1946)	25 elementary, 2 yr. experience	Supervisor ranking	-.01
Jones (1946)	50 high school	Principal rating	.03
	30 high school	Pupil gain	-.23
	13 English	Pupil gain	-.58
Lins (1946)	57 high school women, 1 yr. experience	Composite supervisor rating	.05
	50 high school women, 1 yr. experience	Pupil evaluation	-.34
	17 high school women, 1 yr. experience	Pupil gain	.23

Socioeconomic Status, Sex, and Marital Status Versus Instructor Effectiveness

Socioeconomic Status of Instructor

In 1930 Ullman (339), in an attempt to predict teaching success, among other measures used the Sims Score Card to determine socioeconomic status of 116 junior and senior high school teachers with one semester experience. Near zero coefficients resulted when socioeconomic status scores were correlated with social intelligence, general intelligence, knowledge of principles of teaching, knowledge of aims of secondary education, self-rating, academic marks, education marks, major subject marks, and practice teaching rating. In the case of teaching interest, as measured by the Strong Interest Blank, a coefficient of -.25 was obtained.

This negative relationship appears reasonable considering the low salaries of teachers and the opportunity for individuals of high socioeconomic status to enter professions requiring more costly preparation, but there is no reason why economic status should be related to the other variables. The correlation between socioeconomic status and rated success in the field was .19. Any such low positive coefficient may mean only that supervisors are influenced somewhat by the socioeconomic standing of their teachers. It could mean, too, that persons from the higher socioeconomic group do make better teachers because of greater social poise.

Kriner (182), in 1931, made a study of 131 best and 131 poorest teachers within a school system as judged by superintendents. He found that high school teachers who came from a rural area and whose fathers were farmers and elementary school teachers who came from urban communities and whose fathers were businessmen had the best chance for teacher success. Either type of teacher, especially the elementary, was handicapped if their fathers were artisans and especially handicapped if their fathers were laborers. To enter the teaching profession because of financial reasons or compulsion predicted substantially against teaching success. Size of family affected teacher success probably as a by-product of financial reasons. Travel and past illness had little if any relationship to teacher success. Kriner's results are probably not specific with teachers. They may simply be demonstrating the truism that those from the higher status groups have greater probabilities of success in life than those less fortunate.

Phillips (256) secured ratings by superintendents and principals of 173 elementary and junior high school teachers. He also administered the Sims Socio-Economic Scale to the same group. The resulting correlation coefficient between these measures was .05. When the ratings were converted to sigma scores, Phillips reports a correlation of .22 for the entire group and a critical ratio of 3.5 for two groups of 43 teachers each standing at the extremes of teaching ability as rated administratively.

Rolfe (280) computed correlations between achievement in citizenship of 338 seventh and eighth grade pupils from one- and two-room rural schools and various measures of their 47 teachers. He reported a correlation coefficient of -.15 between the teachers' Sims Socio-Economic Status scores and pupils gains.

The results obtained with the Sims Socio-Economic Scale, like those found with the Cooperative General Culture Test, seem to provide little incentive for further research in this area.

With the exception of Rolfe's (280) study the criterion used in these studies was supervisory ratings, which are often negatively correlated with student gain. It is possible that with other criteria and with other hypotheses involving socioeconomic status of teachers research of more probable productiveness might be undertaken. Socioeconomic status of the teacher is probably not of significance in itself but only as it might be reflected in various "psychological" dimensions of teachers. For instance, if extreme upward social motility has characterized a given teacher and this motility has resulted in insecurity and anxiety on the teacher's part, this might in turn be reflected in the teacher's pattern of classroom behavior or in the adjustments the teacher makes to administrative personnel, fellow teachers, and pupils. Instead of looking for people who have exhibited this social motility, or who possess a certain socioeconomic status, investigation might be directed toward the manifest degree of anxiety or insecurity.

Sex of Instructor as Related to Instructor Effectiveness

In Table 29 are summarized the ten available studies in which sex of instructors was related to instructor effectiveness. It will be noted that criteria of effectiveness employed included student ratings, student designation of best teacher, average class marks, administrative ratings, and success or failure on the job. One investigator used three criteria: pupil gain, pupil ratings, and administrative ratings. Six of these studies appear to favor women, three show no differences between effectiveness of men and women, and two studies favor men. In studies conducted prior to 1940, in no instance apparently was the significance of the obtained difference between teaching effectiveness of men and women teachers tested. In the four later investigations significance was determined but in only one study, that of Cheydleur (75), was a significant difference found, a critical ratio of 6.6 being reported in favor of women instructors.

As indicated by the foregoing studies the question as to whether or not women teachers are superior to men teachers has been considered for some years. The problem may not be merely one of academic interest but may have practical or economic implications for some school and college administrations. No particular differences have been shown when the relative effectiveness of men and women teachers has been compared. In view of the results found, it may well be that consideration should be given to assessing the effectiveness of women instructors in Air Force technical schools. In case of full scale mobilization women, both civilian and WAF, would seem to offer an invaluable potential source of instructional personnel. Employment of greater numbers of women instructors than at present would release like numbers of technical specialists who would then be available for combat support in their specialty.

The Relation of Marital Status to Instructor Effectiveness

While in some parts of the country there has been considerable opposition, generally for economic reasons, to the holding of teaching positions by married women, there appears to be little evidence that married teachers are in any way inferior to unmarried teachers. The reviewers found only three investigators who had made any objective study of the question. In 1934 Peters (253) conducted a rather comprehensive study

Table 29

Sex of Instructor as Related to Instructor Effectiveness

Investigator	Teacher sample	Measure of effectiveness	Differences in favor of Men	Differences in favor of Women	Differences in favor of Neither
Boyce (1912)	343 high school	Administrator ratings			X
Nanninga (1924)	336 men, 896 high school women	(79 superintendents' reports of failure)		X	
	108 men, 2179 elementary women			X	
Birkelo (1929)	(Number unreported)	Percentage of 614 college students designating as best pre-college teachers		X	
Herda (1935)	13 men, 14 women	Pupil rating			X
	(Number unreported)	Pupil gain		X[a]	
	None	67 superintendents opinions			X
Heilman & Armentrout (1936)	31 men, 15 college women	Mean student ratings		X	
Engelhart & Tucker (1936)	(Number unreported)	Percentage of high school students designating as best and poorest teacher ever had		X	

---

[a] Women teachers excelled men in 13 subject fields; men excelled women in 7 subject fields.

[b] Critical ratio = 6.6.

Table 29 (Cont.)

Investigator	Teacher sample	Measure of effectiveness	Differences in favor of Men	 Women	 Neither
Davenport (1944)	51 high school	Student ratings or rankings			X
Cheydleur (1945)	61 men & 9 women, college French	Average class marks		X[b]	
Nemec (1946)	74 men, 191 women probationary longer than 2 yrs.	Teaching certificate finally granted	X		
Cooper & Lewis (1951)	153 student	Student ratings	X		
	72 high school		X		

---

[a] Women teachers excelled men in 15 subject fields; men excelled women in 7 subject fields.

[b] Critical ratio = 6.6.

of the status of the married woman teacher. He matched according to age, education, teaching situation, and so on, 110 married with 110 single elementary school teachers and compared the gain of 2195 pupils of the former group with that of 2250 pupils of the latter group. Supervisory ratings (made by superintendents or principals) were obtained for 1123 married teachers and 1123 single teachers matched on the same variables as above. Differences in achievement and mental growth of pupils of the married women teachers as compared with the single teachers as shown by scores on the Otis Classification Test Parts I (achievement) and II (mental growth) were .86 ± .29 and .60 ± .23, respectively. These differences in favor of the pupils of the married teachers were just under three times the probable error of the differences or on the border line of being significant. Differences in supervisory ratings of married and unmarried teachers were too small to be significant.

In 1951 Ryans (286) compared 99 single women with 107 married women third and fourth grade teachers with respect to ratings made by trained observers. Dimensions observed included 20 items relating to directly observable teacher behavior and 6 items referring to pupil behavior. Comparison of mean criterion scores with respect to marital status revealed no differences that were significant at or near the .05 level of confidence. When the relation of marital status to pupil behavior alone was studied for the 206 teachers, a coefficient of mean square contingency of .11 was obtained.

The Relation of Teaching Aptitude, Attitude Toward Teaching, And Interest to Instructor Effectiveness

Teaching Aptitude versus Instructor Effectiveness

The results of the ten investigators using several measures designed to predict teaching ability show great disparity. In Table 30 entries have been arranged according to teaching aptitude test instead of chronologically in order to improve comparability of studies. As will be seen from Table 30, correlation coefficients between various criteria of effectiveness and the Knight aptitude test ranged from -.10 to .78, the largest being reported by Cooke using nine teacher subjects. The Morris Trait Index-L test, apparently devised to indicate leadership aspects of teaching aptitude, gave correlation coefficients between scores on this test and various criteria of teaching competence from -.17 to .23. In the case of the Coxe-Orleans Aptitude Test the range of coefficients with various criteria of teaching efficiency was -.32 to .51. Dodd (100) suggests that the Coxe-Orleans test measures qualities related to general scholarship rather than to teaching success as revealed by supervisors' ratings of practice teaching. The range for the Stanford aptitude test was -.15 to .14. For the George Washington University Aptitude Test a coefficient of -.19 was reported by Seagoe (299).

Table 30

Relation of Scores on Measures of Teaching Aptitude to Teaching Effectiveness

Investigator	Teacher sample	Measure of effectiveness	Correlation
			Knight Aptitude-Elementary
Knight (1922)	33 elementary	Fellow teacher rating	.45
	7 high school	Fellow teacher rating	.15
	33 elementary	Supervisor rating	.77
	7 high school	Supervisor rating	.00
Tiegs (1928)	25 elementary, 1 sem. experience	Supervisor rating	.02
Bathurst (1929)	(Number unreported) elementary	Administrator rating	.50
Barr, et al. (1935)	66 elementary	Pupil gain A.Q.	-.01
		Pupil gain (composite A.Q. & raw gain)	-.10
		Superintendent rating (composite 7 scales)	.28
Cooke (1937)	27, 18, 9 elementary & high school	Self-rating	.21, .22, .38
		Supervisor rating	.32, .12, .78
			Morris Trait Index-L
Barr, et al. (1935)	66 elementary	Pupil gain A.Q.	-.11
		Pupil gain (composite A.Q. & raw gain)	-.04
		Superintendent rating (composite 7 scales)	.08
Phillips (1935)	173 elementary & jr. high school	Superintendent & principal rating	.20
		Sigma score rating	.23
Rolfe (1945)	47 elementary, 1- & 2-room	Pupil gain	-.17
Rostker (1945)	28 elementary, rural[a]	Pupil gain (social studies)	.20
Seagoe (1946)	25 elementary, 2 yr. experience	Administrator rating	.00
			Coxe-Orleans Aptitude
Coxe & Cornell (1933)	500 (approx.) elementary, 1 yr. experience	Supervisor rating	-.03
	400 (approx.) elementary, 2 yr. experience	Supervisor rating	.03
	112 elementary, 2 yr. experience	Composite observer rating	.08
Phillips (1935)	173 elementary & jr. high school	Average superintendent rating	.16
		Sigma score rating	.28
Cooke (1937)	9-48 elementary & high school	Self-rating	-.32 to .04
		Supervisor rating	-.12 to .51
Seagoe (1946)	25 elementary, 2 yr. experience	Supervisor ranking	.01
			Stanford Aptitude (3 subtests)
Rostker (1945)	28 elementary, rural[a]	Pupil gain	.02, .04, .10
Rolfe (1945)	47 elementary, 1- & 2-room	Pupil gain	.15, -.13, .08
Seagoe (1946)	25 elementary, 2 yr. experience	Supervisor ranking	.02, .04, .14
			George Washington University Aptitude
Seagoe (1946)	25 elementary, 2 yr. experience	Supervisor ranking	-.19

---

[a] Exclusive of 1- and 2-room schools.

In 1952 Jarecke (165) made an initial report of a teaching judgment test he had devised which follows a somewhat different pattern from other tests of this type. Jarecke's instrument is a situational test of a forced-choice ranking type. A list of problem situations typical in the daily life of a teacher is presented. There are five alternate solutions offered for each situation. Solutions are to be ranked in order of favorableness. All solutions are of the nonoptimum or poor type on the theory that good teachers could discriminate between varying degrees of poor alternatives while poor teachers would tend to rank higher the one they themselves might employ. Jarecke reports very high correlations (.68 to .93) when scores on the teaching judgment test were correlated with various criteria of teaching effectiveness. Unfortunately, however, these reported correlations are spuriously high because the population on which they were obtained included the population on which the scoring key was based, thus making it difficult to evaluate the test on the basis of present data.

At first glance it might appear informative to examine the factors that have been considered worth including in tests of teaching aptitude together with the underlying rationale and implicit hypotheses. The reviewers are of the opinion, however, that rationale or hypotheses or the methods used to implement them have been inadequate. If one knew what kinds of things were important to instructor effectiveness and were able to construct devices for measuring both the instructors' knowledge of these things and the probability of their shaping their behavior in accordance with them in an instructional situation, the use of aptitude tests would seem to be a reasonable approach.

It may be that there is a specific aptitude for teaching which is related to effectiveness of teacher performance. Data thus far available, however, either fail to establish the existence of any such aptitude with any degree of certainty or indicate that the tests used were inappropriate to its measurement.

Teaching Attitude versus Instructor Effectiveness

Attitude toward teachers and teaching, as indicated by the Yeager Scale devised for its measurement, appears to bear a small but positive relationship to teacher success measured in terms of pupil gains. Rolfe (280) administered a battery of tests to 47 rural teachers. He reported a correlation coefficient of .22 between pupil achievement in citizenship and teacher scores on the Yeager Scale. He also found a coefficient of .38 between this success criterion and teachers' scores on the Hartmann Social Attitude Test. With 28 teachers as subjects, Rostker (282) reported a coefficient of .45 between teachers' Yeager scores and measurable changes produced in their pupils in social studies. LaDuke (188), who correlated scores of 31 rural teachers on the Yeager test with "objective" tests of pupil gain in attention, appreciation, information, interest, and a composite of these, found coefficients ranging from zero to .20.

Interest in Teaching versus Instructor Effectiveness

Operationally, interest in teaching may be quite different from attitude toward teachers or teaching. That an effective teacher should be interested in teaching would appear to be so obvious as to be axiomatic. A few investigators have attempted to show that among successful teachers interest in teaching developed during the teachers' secondary school period or before. In the majority of investigations, however, interest in teaching was measured by interest test scores which indicate similarity of interests of teachers and persons undergoing the interest test. The results of these studies are shown in Table 31.

As will be seen from Table 31 those correlations resulting from the use of the Strong interest test or modifications of it and the test used by Coxe and Cornell (87) all tend to cluster around zero. The Link Activities and Interest Inventory on the other hand shows such inconsistencies in the light of the rather sparse data available as to render it also of somewhat doubtful value.

The Kriner (182) study which produced such high correlations was based on recall by the teachers as to their interests when they were in high school. Obviously there is no way of keeping such opinions free from the influence of later experience of success or failure, thus making the correlations obtained practically meaningless. The Lins (203) investigation, on the other hand, was a follow-up study. Students listed their choices as to occupations when they first entered college and these choices were correlated against rating received some years later.

In 1952 Ringness (277) reports a study in which he attempted: (1) to discover, if possible, any common factors that may underlie the reasons given by undergraduates for the choice of teaching as a profession; (2) to determine whether the answers given to essentially the same questions in two different types of testing devices reveal comparable data; and (3) to investigate the relationship between the reasons given for choice of profession and subsequent teaching success as measured by criteria of efficiency and acceptability. A paired-comparison and a ranking questionnaire were used to determine the reasons for choice of teaching as a career. Data were analyzed by the centroid method of factor analysis to find the common factors. Sixty-three men and 37 women student teachers comprised the sample used in Parts One and Two of the study, and 16 men and 18 women with one-year experience were used in the last part of the study. Criterion of teaching success was an "acceptability" rating by the superintendent. This was an over-all rating made after an interview of the superintendent by the investigator in which questions were asked which related not only to teaching efficiency but also to personality traits of many kinds. In the factor analysis study factors identified as interests in working conditions, in people, in security, and in subject matter area to be taught seemed to be generally emphasized. Desire for professional advancement did not appear to be a general characteristic of the factor

Table 31

Relation of Interest Test Scores to Teaching Effectiveness

Investigator	Teacher sample	Measure of effectiveness	Test	Correlation
Ullman (1930)	116 high school, 1 sem. experience	Supervisor rating	Cowdery & Strong	.02
Kriner (1931)	76 poor & 74 best elementary	Superintendent rating	Interest in teaching while in high school	.74[a] .55[b]
	54 poor & 56 best high school		Rather teach now than do anything else	.87[a] .92[b]
			Interest in teaching as a career while in high school	.79[a] .69[b]
			No interest in teaching while in high school	-.64[a] -.62[b]
Coxe & Cornell (1933)	500 (approx.) elementary, 1 yr. experience	Supervisor rating	Coxe & Cornell	-.01
	400 (approx.) elementary, 2 yr. experience	Supervisor rating	Coxe & Cornell	.07
	112 elementary, 2 yr. experience	Composite 2 observer ratings	Coxe & Cornell	.10
Barr, *et al*. (1935)	66 elementary	Pupil gain A.Q.	Strong interest	.03
		Pupil gain (composite A.Q. & raw gain)	Strong interest	.06
		Superintendent rating (composite 7 scales)	Strong interest	-.11
Phillips (1935)	173 elementary & jr. high school	Average superintendent & principal rating	Phillips & Manson	.16
		Sigma score rating	Phillips & Manson	.10
Sandiford, *et al*. (1937)	Top and bottom thirds 420 (approx.) student	Practice teaching grade	Expect to teach a lifetime	3.5 (Critical ratio)
Seagoe (1946)	25 elementary, 2 yr. experience	Supervisor ranking	Strong interest	-.08
Jones (1946)	49 high school	Supervisor ranking	Link Activities & Interest	.36
	27 high school	Pupil gain	Link Activities & Interest	.07
	12 English	Pupil gain	Link Activities & Interest	-.34
Lins (1946)	41 high school women, 1 yr. experience	Supervisor rankings	Teaching listed as first choice of occupation when teacher entered college	.27
Espenschade (1948)	46 physical education, 1 yr. experience	Supervisor rating	Strong interest	.07

---

[a] Pearson cos π *r* formula, elementary teacher sample.

[b] Pearson cos π *r* formula, high school teacher sample.

structure, nor did desire for service to society or prestige and respect of the profession. Factors bearing similar labels were found in analyzing the results for men and women. However, these factors were only broadly alike and had somewhat different arrangements and loadings of the variables. An interest in "security," for example, as interpreted from the men's data is not precisely that interpreted from the women's data. Correlations between reasons for choice of teaching as a profession and acceptability ratings differed slightly between the men and women subjects. Items which had a correlation of .30 or higher, in either the paired-comparison or ranking questionnaire, for the women were: "relatively good financial reward," "ease of getting a position," "clean, attractive physical surroundings," "short working hours," "frequent vacations," and "environment of interesting co-workers." Items which had a correlation of .30 or higher for the men were: "security against job loss and layoffs," "clean, attractive physical surroundings," "opportunity for professional advancement," "opportunity to serve society," "ease of getting a position," and "opportunity to pursue a favorite interest." Multiple-correlation coefficients between acceptability ratings and raw scores in the men's paired-comparison questionnaire were .64, and for the women's questionnaire .44. Multiple correlation coefficients between acceptability ratings and raw scores for the ranking questionnaire were .76 for men and .78 for women. It appears to the reviewers that Ringness may have gone somewhat further in his interpretation of his data than the size of his N's justifies.

The Relation of Voice and Speech Characteristics To Instructor Effectiveness

Shannon (303), in 1928, reported that the teacher's voice was placed eleventh in order of importance among qualities listed by 3317 high school pupils and ninth in importance by 107 university students. One hundred twenty-four critic teachers placed voice second among personal and social traits considered essential to effectiveness that were found to be weak in student teachers under their direction. Voice did not appear among the 15 most important qualities mentioned by 97 supervisors.

In 1951 Richey and Fox (274) had 1883 high school boys and 2022 high school girls in Indiana check characteristics that pertained to their best-liked and least-liked teachers. Among characteristics of the best-liked teachers, the item, "had a pleasant speaking voice," was marked by 76% of the boys and by 84% of the girls. Of the characteristics of the least-liked teachers, "had bad speaking voice" was designated by 39% of the boys and by 37% of the girls.

In other investigations discussed under the section on Opinion Studies voice was mentioned among the ten most important teaching characteristics in eight studies of high school pupils, nine studies of college students, and two studies of administrative groups. Voice was not included among the first ten traits in opinion studies of two grade school groups, four studies of high school groups, seven college student studies, one study of administrative opinion, and two opinion studies of teachers themselves.

In 1929 Barr (16) studied the characteristic differences in the teaching performance of 47 good and 47 poor teachers of the social studies. Twelve of the good and 17 of the poor teachers were listed as having good voices. Twenty-five good teachers and 7 poor teachers showed "conversational manner." A repetition of the study with another group of teachers produced similar results.

In 1941 Baxter (25) in an investigation of teacher-pupil relationships reported results when 42 teachers were studied by two observers. Voice and manner of effective teachers were said to be original and intriguing while noneffective teachers showed voice and manner that were prosaic and colorless.

In 1943 Henrikson (150) made some comparisons of ratings of voice and teaching ability. Teachers were selected at random from the files of a placement bureau. Results are shown in Table 32.

Table 32

Relation of Ratings of Voice and Teaching Ability

Variables	No. of cases	Correlation coefficient
Voice rated by supervisor of practice teaching vs. voice rated by school supervisor	433	.20
Teaching ability rated by school supervisor vs. voice rated by practice teaching supervisor	433	.20
Teaching ability rated by practice teaching grade vs. voice rated by supervisor	432	.27
Teaching ability vs. voice rated by same judge:		
Training school supervisor	434	.62
Public school supervisor	580	.58

The last two correlation coefficients in Table 32 appear to be a rather neat demonstration of the inability of judges to separate supposedly different characteristics of individuals, viz., teaching ability and voice. Other good examples of the same inability on the part of raters can be seen in the investigations of Martin (218) and Henrikson (151). In 1944 when Martin correlated superintendents' ratings of 123 teachers after their first year of teaching with the same superintendents' evaluation of voice and mechanics of speech, she obtained a correlation coefficient of .58. In a later study in 1949 Henrikson (151) investigated relations between personality, speech characteristics (voice, pitch, rate, quality), and teaching effectiveness of college teachers as rated on a five-point scale by 150 college students enrolled in a speech course. He reported coefficients of contingency ranging from .42 to .66 and chi-square values showing significant relationships between various qualities of instructors as determined by the student ratings.

From the studies reviewed above it appears, in general, that the quality of the teacher's voice is not considered too important by school administrators, teachers, and students. Halo effect or "logical error," which so often has been found a contaminating factor in ratings, also appeared to be present to a large extent in these studies.

A study, made by McCoard (210) in 1944 on speech factors as related to teaching efficiency, appears somewhat more promising. Speech effectiveness of 40 teachers in one-room schools was measured by having 22 speech teachers rate each teacher on a seven-point scale on each of 14 speech factors. Recordings were made while each teacher read standardized material for three minutes and also spoke for three minutes on an assigned topic. A special pronunciation test was also administered. Correlations were obtained between the gains of 338 seventh and eighth grade pupils in a citizenship test and their teachers' speech scores. In the reading experiment 12 of the 14 ratings on speech factors and the total speech score were significantly correlated with student gains at the .01 level, and the correlations of the other two speech factors with student gains were significant at the .05 level of confidence. The coefficients ranged from .34 to .46. In the speaking experiment two speech factors, variation in pitch and variation in quality, had correlations with student gains that were significant at the .01 level. Eight speech factors and the total were significantly related to gains at the .05 level of confidence.

Correlations obtained between a composite of effectiveness ratings by supervisors and reading scores were all significant at the .05 level and all but two were significant at the .01 level of confidence. The correlation between total speech scores (reading and speaking combined) and supervisors' ratings was .49. Intercorrelations among various speech factors (pitch, quality, volume, rate, phrasing, distinctness, etc.) centered around .90 which led the author to conclude that even with trained judges an indication of general speech ability based on a single factor will give as good results as a total of judgments on several factors. McCoard reported correlation coefficients between pronunciation test scores and other teacher measures as follows: pupil gain .02, supervisors' ratings .40, total reading score .49, total speaking score .40.

In 1950 Huckleberry (159) investigated the possible relationship of speech to student teaching. He also attempted to develop means of identifying significant speech qualities of student teachers and observed the effect of improvements of speech on student teaching competency. Three speech teachers rated recordings of 54 volunteer subjects (24 in the experimental group and 30 in the control group) in terms of articulation, pronunciation, voice quality, voice pitch, inflection, rate, rhythm, and conviction. Huckleberry concluded that positive change in student teaching proficiency, as observed by critic teachers, was directly associated with positive change in rated speech proficiency. The reviewers compared the correlation coefficients of his experimental and control groups, however, and found the differences were not statistically significant.

While research on voice and speech characteristics tends to be somewhat scanty, this area appears promising for research in the Air Force technical school situation. It is possible that voice, apart from other variables, plays an important part in supervisors ratings. It may be, too, that speech characteristics constitute a crucial instructor variable, that in addition to certain subject-matter knowledge or other prerequisites, the competent instructor is the one whose voice appeals to his class. It may be, on the other hand, that "actions speak louder than words," that the instructor who "knows his stuff" and is able to demonstrate his knowledge has little need for words. A potentially fruitful research approach to this problem might be first, to determine the extent to which student gains in Air Force schools are related to the instructors' oral presentation; and second, to determine whether or not this ability can be measured prior to selection for the instructor assignment.

The Photograph as a Predictor of Instructor Effectiveness

Many school administrators require a photograph of the applicant to accompany letters of application for teaching positions. In order to determine the validity of this alleged aid to selection, Tiegs (333), in 1928, evaluated photographs as a means for teacher selection. He reported that rankings by five judges of teaching effectiveness of 25 elementary school teachers on the basis of photographs gave rise to intercorrelations among them ranging from .00 to .50. Official ratings of the 25 teachers given by superintendents, after the rating forms had been checked by principal and general supervisor, when compared with rankings by photograph produced a correlation coefficient of -.08.

Johns and Worcester (169), in 1930, also attempted to submit the photograph to an experimental check. In their study 6 faculty members of a teachers college ranked 6 men school superintendents or principals, 6 high school, 6 elementary school, and 6 kindergarten and primary women teachers on teaching effectiveness. Photographs of these 24 administrators or teachers were then mailed to 148 judges: 61 superintendents, 38 school board secretaries, and 49 placement bureau secretaries. The judges were asked to rank members of each group from the photographs as to their desirability as teachers. The results showed every photograph assigned every rank from one to six by every class of judge. Correlations between composite rankings of judges of photographs and faculty committee rankings were: for superintendents and principals, -.10; for high school teachers, .14; for elementary school teachers, -.01; and for kindergarten primary teachers, .37. No one judge of photographs in the whole 148 agreed with the faculty committee ratings for any of the groups ranked.

Statistical Analyses of Instructor Abilities

Nine studies (12, 70, 142, 149, 189, 277, 295, 287, 288, 289, 314) report results of factor analyses of data from presumed measures of teaching abilities. In 1932 Butsch (70) by means of a tetrad difference analysis found a general factor among the intercorrelations of judgments of teacher traits. In 1943 Smalzried and Remmers (314) applied the Thurstone method of factor analysis to student ratings of 40 practice teachers on the Purdue Rating Scale for Instructors. Two factors emerged which they designated "empathy" and "professional maturity." Items which had greater saturation of "empathy" were fairness in grading, personal appearance, sympathetic attitude toward students, and liberal and progressive attitude. The items with the greater loading for "professional maturity" were self-reliance, confidence, and presentation of subject matter. The other items of the scale show lower and more nearly equal saturation with both basic factors.

Hellfritzsch (149), in 1945, reported a factor analysis of some 27 teacher variables using data from the Rostker (282) and Rolfe (280) studies. He concluded that four independent primary teacher abilities satisfactorily explain the intercorrelations observed between a battery of measures commonly used in investigations of the nature, measurement, and prediction of teaching ability. These he identified as: general knowledge and mental ability; teacher rating scale factor; personal, emotional, and social adjustment; eulogizing attitude toward the teaching profession. The four factors were uncorrelated with each other. Each of the several teacher measures was dependent primarily upon only one of the factors. Hellfritzsch also stated his study revealed that no single teacher measure of those he used could validly be substituted for the actual measurement of pupil growth in evaluating the ability of teachers to teach. Supervisory ratings, he found, were only slightly related to observed pupil growth in social studies and, hence, Hellfritzsch concluded were of doubtful value as a measure of teaching effectiveness conceived in terms of ability to promote pupil growth.

In 1950 Schmid (295) conducted an investigation to determine by means of factor analysis if a few common factors might adequately summarize areas of personality and ability of prospective teachers. Scores were obtained by means of the Washburne Social Adjustment Inventory, Mooney

Problem Check-List, Minnesota Multiphasic Personality Inventory, and personal data from student files with respect to 24 traits for 51 male and 51 female student teachers. The size of the total group tested varied from 80 to 101 for the different variables. Schmid hypothesized that the factor patterns would differ for males as compared to females and ran separate analyses by sex. Unfortunately this reduced the number of individuals represented by each correlation coefficient to such low figures (40 to 51) as to make the results of his analyses highly tentative. Factor analysis of female scores yielded four common factors, identified as "problems in response set," "professional maturity," "introversion," and "social adjustment." In general, Schmid says, these factors failed to cut across areas measured by the personality measures he used perhaps indicating that these instruments are measuring different aspects of personality. Factor analysis of the male scores resulted in two common factors, "social and educational adjustment" and a "personality-psychological" factor. The factor pattern of the male students showed a marked discrepancy from that of the female students.

In 1951 Lamke (189), in a factor analysis of personality characteristics as measured by Cattell's 16 Personality Factor Test for 10 good and 8 poor high school teachers with one year's experience, found that responses of good and poor teachers did not fall into two well-defined and characteristic patterns. There was some indication that some good teachers differed from some of the poor teachers on the responses associated with Cattell's source traits F (surgency vs. desurgency or anxious agitated melancholy), H (adventurous cyclothemia vs. withdrawn schizothemia), and N (sophistication vs. simplicity). The reviewers are inclined to doubt the significance both statistical and practical of factor analytic studies based on 18 cases.

Ryans as part of the "Teacher Characteristic Study" has made a factor analysis of trained observer ratings of elementary and secondary teachers on a classroom observation scale containing 20 items referring to teacher behaviors and 6 referring to pupil behavior. Results of this study have been published in a number of different references (287, 288, 289). A detailed account of the factors found is given in the section on Objective Observation of Instructor Performance.

In 1951 Hampton (142) published the results of a factor analysis of supervisory ratings of elementary teachers. Two different scales, a paired-comparison scale and a graphic rating scale, were used. Hampton concluded that a general factor did not account for the intercorrelations of the ratings on either instrument. Furthermore, that a greater number of factors was needed, namely six as compared with three, to account for the intercorrelations of the same traits on the paired-comparison instrument than was needed to account for the intercorrelations of the ratings on the graphic scale.

In 1952 Bach (12) used the factor analysis approach in a study of the relationship of critic teacher ratings as student teachers and supervisory ratings of the same subjects after they had had actual teaching experience. Bach found four factors for each of the two ratings, but only two of these appeared to be similar.

In 1952 Ringness (277) factor analyzed data concerning reasons given by teachers for choice of teaching as a career. This material is discussed in detail in the section on The Relation of Teaching Aptitude Attitude Toward Teaching and Interest to Instructor Effectiveness.

Two investigators (220, 285) have reported results of item analyses of instructor traits, one in terms of student change and the other in terms of principals' assessments.

In 1940 Mathews (220) made an item analysis of measures of teaching ability in relation to student change. By means of a battery of tests he derived a composite index of the changes produced in seventh and eighth grade pupils by 57 rural school teachers of social studies. The teachers were given a battery of 11 psychological, subject matter, and adjustment tests. Of the 1675 items in all tests given the teachers only 68 items, or slightly over 4%, possessed statistical significance in terms of pupil change. Mathews concludes that the findings cast serious doubt on the validity of the tests studied as measures of teaching ability when pupil change is used as a criterion.

Ryans (285), in 1951, applied analyses of internal consistency and external validation procedures to test items measuring the professional information of 192 elementary and 165 secondary teachers with one or more years' experience. He used three teacher measures: (a) scores on the General Principles and Methods of Teaching Test of the 1949 National Teacher Examination battery; (b) principals' assessments by means of an observation blank of teacher behavior in terms of pupil behavior, teacher personal-social behavior in the classroom, and teacher behavior indicative of intellectual and educational background; (c) principals' general evaluation of teachers' over-all effectiveness on a graphic rating scale. The two principals' ratings produced an intercorrelation coefficient of .83 for both elementary and secondary teacher groups which might be expected because of the common factors involved. Upper and lower 27% of teachers were segregated and analyses of the three measures and item discrimination indexes for the teachers' test were computed for these groups. The General Principles and Methods of Teaching Test, Ryans concluded, appeared to be made up of items that functioned satisfactorily from the standpoint of internal consistency. However, when the test items were analyzed against either of the principals' ratings less than 20% of the 45 items discriminated significantly at the .05 level or better between high and low elementary teachers. Only 5% of the items discriminated between high and low secondary teachers. Ryans attributes these somewhat unsatisfactory results to ". . . the doubtful validity and reliability of the assessments upon which the external criteria were based, the low reliability of individual items, and the fact that understanding of educational concepts comprises only one segment of over-all teaching effectiveness. . ."

In the opinion of the reviewers all the studies of the foregoing types so far reported suffer from inadequacies of criteria, tests, or numbers of cases. It still seems possible that a more adequately designed study might yield results of considerable basic importance to the solution of problems of evaluating and selecting instructors.

Opinion Studies of the Personality Characteristics Of Effective and Ineffective Instructors

For over fifty years attempts have been made to identify the personality characteristics of successful and unsuccessful teachers by making lists of traits based on opinions. In most cases these lists have been made up of subjectively estimated characteristics of such a vague, general nature as to render any precise measurement of them impossible. One of the earliest studies of this kind was that made by Kratz (181) in 1896. When 2411 pupils were asked to indicate the characteristics of their best teachers, the factors most frequently mentioned were: helped in studies, personal appearance, good, kind, pleasant, happy, jolly, patient, polite, neat. In 1929 Charters and Waples (74) collected some 2800 teacher traits as reported by 27 teachers, 14 parents, 10 pupils, 3 teacher agency executives, and 2 professors of education. It might be thought that this exhaustive and comprehensive list would be the list to end all lists. However, more papers using this approach have appeared since 1929 than ever appeared before that date.

In the search for traits, qualities, and characteristics of the successful teacher, almost no stone has been left unturned. Table 33, lists all available studies categorized according to the group from whom opinions were solicited. The studies are arranged in chronological order under each category.

Several of the opinion studies that are somewhat interesting because of the novelty of the approach employed, the date of the study, or the magnitude of the effort involved will be briefly reviewed.

In 1900 Bell (27), in a study of the teacher's influence, reporte results of a questionnaire completed by 543 men and 488 women normal school students. In indicating characteristics of those teachers that were most helpful the students' answers fell into four groups: (1) moral influence; (2) personal interest, kindness, encouragement, sympathy; (3) intellectual influence; (4) self reliance. Almost all students indicated that they had had a teacher whom they positively disliked or hated. The disliked teachers were reported to have a malevolent attitude, either active or passive, resulting in such behavior as unjust punishment, sarcasm, insult, and ridicule.

Shannon (303), in 1928, made a most comprehensive investigation of opinions of the personal and social traits of successful and unsuccessful

Table 33

Opinion Studies of Traits, Qualities, and Characteristics of Successful Teachers

Study	Year
Elementary school pupils' opinions	
Frasier (120)	1927
Messler (226)	1932
Jersild (168)	1940[a]
Witty (356)	1947
Witty (357)	1948[a]
Junior high school pupils' opinions	
Schaffle (293)	1931
Messler (226)	1932
Davies (93)	1933
Jersild (168)	1940
Daniel (91)	1944
Witty (356)	1947
Witty (357)	1948
Senior high school students' opinions	
Book (43)	1905
Bird (35)	1917
Graves (134)	1925
Davis (95)	1926
Ryle (290)	1928
Shannon (303)	1928
Jordan (173)	1929
Shaffle (293)	1931
Bowman (47)	1934
Krous (186)	1935
Hart (145)	1936
Engelhart & Tucker (108)	1936
Jersild (168)	1940
Albert (1)	1941
Daniel (91)	1944
Smith (317)	1945
Witty (356)	1947
Witty (357)	1948
Irwin (162)	1949
Leipold (199)	1949
Richey & Fox (274)	1951
Normal school students' and student Teachers' opinions	
Bell (27)	1900
Bird (35)	1917
Newmark (245)	1929
MacDonald (211)	1931
Malan (215)	1931
Krous (186)	1935
College undergraduates' opinions	
Dolch (103)	1920
Davis (95)	1926
Breed (53)	1927
Champlin (73)	1928
Copper (84)	1928
Shannon (303)	1928
Birkelo (36)	1929
Mead (221)	1929
Clinton (78)	1930
Ludeman (206)	1931
Greene (135)	1933
Bousfield (45)	1940
Lamson (190)	1942
Haggard (141)	1943
Smith (316)	1944
Goodharts (131)	1948
Bradley (50)	1950
Riley, *et al.* (276)	1950
Opinions of teachers in service	
Breed (53)	1927
Frasier (120)	1927
Charters & Waples (74)	1929
Jordan (173)	1929
Cattell (72)	1931
Malan (215)	1931
Daniel (91)	1944
Critic teachers' opinions	
Shannon (303)	1928
Shannon (306)	1941
Opinions of teacher training specialists	
Charters & Waples (74)	1929
Cattell (72)	1931
Tostlebe (334)	1937
Opinions of administrators:	
Supervisors	
Shannon (303)	1928
Charters & Waples (74)	1929
Jordan (173)	1929
Albert (1)	1941
Shannon (306)	1941
Daniel (91)	1944
Superintendents	
Anderson (5)	1917
Frasier (120)	1927
Bell (28)	1932
Tostlebe (334)	1937
Shannon (306)	1941
Daniel (91)	1944
Principals	
Hill (156)	1929
Shannon (306)	1941
Daniel (91)	1944
Directors of education, inspectors	
Cattell (72)	1931
College executives	
Kelly & Anderson (174)	1929
Reed (263)	1935
Opinions of parents	
Kephart (175)	1922
Charters & Waples (74)	1929
Jordan (173)	1929
Daniel (91)	1944
Opinions of executives of three teacher Employment agencies	
Charters & Waples (74)	1929

---

[a] Including Grade 1.

secondary school teachers. He interviewed 97 "selected" supervisors; he had 3317 high school pupils and 107 university students list good and bad qualities of teachers; and he asked 124 critic teachers to list personal and social traits found to be weak in student teachers under their direction. Shannon also studied the problem by making analyses of traits used on rating scales, recommendation procedures, reasons for teacher failure, traits considered in certification and codes of professional ethics for teachers. Among teacher traits Shannon found to be considered most important were such qualities as stimulative power, forcefulness, sympathy affability, self-control, and fairness.

In 1929 Jordan (173), in a study of personal and social traits as related to high school teaching, used a questionnaire of 46 traits. The 15 traits considered of most importance, the 16 of medium importance, and the 15 of least importance were checked by 150 high school pupils, 120 teachers, 100 supervisors, and 120 school patrons. As an example of the outcome of typical studies of this kind, the 5 most important and the 5 least important traits as listed by the various groups are given in Table 34. The rather remarkable agreement among the four groups studied suggests the probable existence of powerful cultural stereotypes in the region where the study was conducted. This conclusion is emphasized by the comparative lack of importance indicated by other studies of certain factors judged among the most important in Jordan's study.

Table 34

The Five Most and the Five Least Important of 46 Teacher Traits As Ranked by Four Groups of Judges[a]

	Pupils	Teachers	Supervisors	Patrons
Most important trait				
1.	Fair	Intelligent	Tactful	Intelligent
2.	Intelligent	Tactful	Intelligent	Fair
3.	Interesting	Healthy	Fair	Broad-minded
4.	Broad-minded	Broad-minded	Cooperative	Tactful
5.	Cheerful	Cooperative	Healthy	Patient
Least important trait				
42.	Dignified	Trustful	Ready of speech	In touch with life
43.	In touch with life	Willing to lead	Of broad interests	Trustful
44.	Thoughts centered outside of self	Reverent	Thoughts centered outside of self	Proud of profession
45.	Reverent	Modest	Willing to lead	Of broad interests
46.	Proud of profession	Thoughts centered outside of self	Modest	Willing to lead

---

[a] Jordan (173).

In 1929 Klopp (177) gave results obtained by asking summer school pupils in junior and senior high schools to compare 81 practice teachers with an "ideal teacher" on 10 traits. A majority of the pupils rated their student teachers as equal to the ideal teacher on eight of these traits (kindness, neatness, fairness, patience, approachableness, sense of humor, enthusiasm, willingness to help). Percentages for the different traits ranged from 56% to 78%. The majority rated their teachers below the ideal teacher for thoroughness (55%) and discipline (62%).

In 1932 Kyte (187) asked 69 supervisors to analyze their most serious problem teacher. The supervisors rated their unsuccessful teachers on 53 characteristics. Among these deficiencies judged most important were deficiencies in leadership, in influence on pupils' habits, in selection of method, in control of class, and in work responsibility.

In 1936 Engelhart and Tucker (108) asked 224 high school pupils to check a list containing 100 positive traits and their corresponding opposites for the teacher they considered best and also for the one considered the poorest. Of the 100 traits, 46 were found to correlated significantly and positively with quality of teaching. The highest tetrachoric coefficient of correlation was .93, the 46th was .32. Of the 46 traits correlated 25 were .72 or above. The traits showing tetrachoric coefficients of .80 or higher were good judgment .93, clear in explanation .88, respecting others' opinions .86, sincere .83, impartial .83, fair .82, appreciative .80, interested in pupils .80, broad-minded .80.

Tostlebe (334) made an analysis of the relative importance of various training factors to success in the one-room rural school. A check list of 135 items arranged as a four-point scale was marked by 40 specialists in the field of teacher training and 40 county superintendents. Split-half reliability coefficient for the specialists was .85 and for the county superintendents .86. A correlation coefficient of .81 was obtained between the judgments of the 40 specialists and the 40 superintendents. A weighted index was obtained for each of the 135 success factors which were then divided into fourths. The type of success factors which most predominated in the top fourth were those centering about assignments, individual differences, study periods, mastery of fundamentals, unit method of instruction, adjusting programs, teacher's personal self, and the relationships of the teacher to child and parents.

Daniel (91) compiled opinions of 202 superintendents, 267 principals, 29 supervisors, 846 white teachers, 602 Negro teachers, 1659 white eighth grade pupils, 523 Negro eighth grade pupils, 998 white eleventh grade pupils, 378 Negro eleventh grade pupils, 1351 white patrons, and 973 Negro patrons. Each of the above individuals indicated the qualifications of the teacher whom they considered best within their experience. All groups followed remarkably similar patterns giving first place to qualities related to professional interest and competency, followed by personal qualities.

In 1948 Witty (357) listed in order of frequency traits found in 14,000 letters submitted by pupils from Grades 1 to 12 in a contest which required them to describe the teacher who had helped them most. In a second study of 33,000 such letters the list remained substantially the same. The 12 most frequently mentioned traits in order were: cooperative and democratic attitude, kindliness and consideration of the individual, patience, wide interests, personal appearance and pleasing manner, fairness and impartiality, sense of humor, good disposition and consistent behavior, interest in pupils' problems, flexibility, use of recognition and praise, unusual proficiency in teaching.

Undesirable characteristics were also analyzed in the second study. In order of frequency the 12 most often mentioned negative factors were: bad tempered and intolerant, unfair and inclined to have favorites, disinclined to show interest in the pupil and to take time to help him, unreasonable in demands, tendency to be gloomy and unfriendly, sarcastic and inclined to use ridicule, unattractive appearance, impatient and inflexible, tendency to talk excessively, inclined to talk down to pupils, overbearing and conceited, lacking in sense of humor.

In a study made at Brooklyn College and reported by Goodhartz (131) in 1948 and by Riley et al. (276) in 1950, 6681 students at Brooklyn College selected from a 10-item list, 3 qualities which they considered to be of the greatest importance in a teacher in the biological and physical sciences, the social sciences, and the arts. This study has a certain unique value in that it secured opinions concerning teachers of different subject matter and did not assume that all good teachers would have the same qualities regardless of the subject they taught.

In 1949 Irwin and Irwin (162) obtained an appraisal of certain teacher traits by 415 senior high school students by having them list words that might be used in describing good and bad teachers.

Using 694 students from four college classes Bradley (50), in 1950, using an unstructured, open-end questionnaire technique, found that with respect to college teachers and their teaching, students like such factors as "teaching efficiency," "meets students' needs," "puts subject matter across," "facilitates learning." These were mentioned 1649 times, or more than all other factors put together. Similarly, in terms of dislike, the negatives of these factors appeared 1507 times, again more often than all the other negative characteristics combined.

The results of all of this effort in conducting opinion studies of instructor personality characteristics appear to be largely sterile in terms of usability for evaluative or selective purposes. It seems quite possible that anyone who had passed through the average American school system could sit at his desk and devise an "armchair" list of characteristics of the effective as opposed to those of the ineffective teacher that would be quite as useful as any list thus far developed. The trend in present day research in the area of selection and evaluation of personnel is definitely directed away from opinion studies as sources of ideas concerning the requirements of teaching and toward the use of psychological theory and rationale in the development of systematic sets of hypotheses to be tested with objective tests and observational techniques.

Carefully designed opinion studies of personality characteristics of instructors might lead to some understanding of why supervisors' ratings of instructor effectiveness, which are based on opinion, fail to correlate with the student gains criterion. Investigation might also be directed toward the problem of providing sounder bases for supervisor judgment. It is possible that in such studies the use of some of the more recent methodological refinements such as Stephenson's Q-technique or Cattell's R-technique might be productive of more operationally useful results.

It should be pointed out perhaps that mere collection of great masses of data does not necessarily produce a more effective study. Adequate sampling might have eliminated, for example, in the studies of Witty (356, 357), the arduous task of going through 33,000 or even 14,000 letters without sacrifice of any meaningful finding.

Causes of Teacher Failure

In a number of studies attempts have been made to set forth the causes of teacher failure. Several of these (60, 68, 187, 213, 216, 234, 237, 278, 311) merely report summaries of superintendents' reasons for dismissal of unsatisfactory teachers or give superintendents' opinions as to what constitute the chief weaknesses of failing teachers. Other investigators, Andersen (5) and Morrison (231), include reasons for failure as reported by school board members. Mott (236) queried 200 teachers of agriculture, while James (164) canvassed opinions of college freshmen, school administrators, and teachers themselves. School principals were included in Littler's (205) survey. McLaughlin (212) made a case study of 98 effective and 16 ineffective female elementary school teachers.

The first such report available to the reviewers, that of Littler (205), in 1914, mentioned weaknesses in maintaining disciplines in teaching skill, interest, personality, effort, and cooperation as the most important causes of teacher failure. Subsequent studies have more or less reiterated in somewhat varied terms the findings of this earlier report. Poor maintenance of discipline and lack of cooperation tend to be listed among the chief causes of failure or dismissal in most of these studies. Health, educational background, training, age, and knowledge of subject matter, on the other hand, appear to be relatively unimportant factors. These investigations are marked by a complete absence of operational definitions of the terms used, so that any estimate as to the importance of the various factors depends entirely upon the personal likes and dislikes, preconceptions and misconceptions of the judges and upon their individual interpretation of the terms. In none of these studies was any attempt made to observe unsuccessful teachers systematically in order to determine those specific behaviors which differentiate the ineffective from the successful teacher. Another important consideration in evaluating these studies of the causes of teacher failure is that the stated causes may have been concocted after the decision to relieve the teacher of further duties had been made.

As in the case of opinions regarding the unsuccessful teacher, many judgments have also been made as to what constitutes good teaching practice. No one knows, however, to what extent manifestly undesirable behavior may be offset by presumably desirable factors. In other words, no one has determined what constitute the allowable instructor idiosyncracies. A potentially fruitful approach to the problems of determining instructor effectiveness might well be the investigation, through objective observation techniques, of behavior characteristics commonly deemed to constitute unsound teaching practices by educational authorities. Then study should be made of the extent to which such pedagogically undesirable behaviors may be present without appreciably reducing the efficiency of an instructor in terms of pupil gain.

Personality Tests of Teachers

Investigations of the relations of personality test scores to measures of teacher success have yielded widely varying results. In Table 35 are summarized results of studies in which attempts have been made to related various personality measures to measures of instructor effectiveness. The material has been grouped according to the personality measure used. It will be noted that correlation coefficients computed between scores obtained on the several sections of the Bernreuter Personality Inventory and various criteria of instructor effectiveness range, for "neurotic tendency" from -.31 to .17, for "self-sufficiency" from -.24 to .20, for "dominance-submission" from .00 to .33, for "extroversion-introversion" from -.14 to .01. Correlation coefficients for the Bernreuter-Flanagan self-confidence scale range from -.38 to .00 and for the Bernreuter-Flanagan sociability scale from -.26 to -.06.

High scores on the Bell Adjustment Inventory and on the Thurstone Personality Schedule are associated with poor adjustment so that negative coefficients with effectiveness might be expected. As reported by various investigators these range from -.04 to -.40. The positive coefficients given in the Gould (133) study probably indicate only that he reversed the direction of his scores so that the results among several sets of variables would have comparable directions. Although the tetrachoric correlation of .52 found by Cooper and Lewis (83) between pupil rating and absence of neurotic sign on the Rorschach is higher than is usually found with supposedly more "dependable" data, the authors point out that extent of overlapping prohibits the use of neurotic signs for individual prediction. An important feature of the Cook and Leeds (80) and Leeds (198) studies was the use of item analysis against the external criterion of teachers designated by their principals as the best and worst in the schools in getting along with children.

Ryans (286), in 1951, as part of the "Teacher Characteristic Study" referred to earlier, studied the relationship of scores on the Thurstone Temperament Schedule for the upper and lower 27% of a group of 275 elementary teachers selected on the basis of composite observer ratings. These ratings had been factor analyzed by the centroid method and yielded five oblique factors which appeared to refer to: (a) pupil participation and teacher open-mindedness; (b) controlled pupil activity and businesslike approach; (c) teacher calm and consistent, liked because "human;" (d) sociability; (e) appearance and attractiveness. (This last factor was not used in the analysis.) Differences for the "vigorous" category of the Thurstone Temperament Schedule were significant at the .01 level for Factor (a); for the "impulsive" category at the .05 level for Factor

Table 35

Relation of Personality Measure to Measures of Instructor Effectiveness

Investigator	Teacher sample	Measure of effectiveness	Correlation					
			Bernreuter Personality Inventory					
			N	S	I	D	FC	FS
Laycock (1934)	80 student	Practice teaching grade	-.21	.05	-.14	.33		
Phillips (1935)	173 elementary & jr. high school	Average superintendent & principal ratings	-.09	.04	.07			
		Superintendent rating (sigma scores)	.14	.01	.21			
Heilman & Armentrout (1936)	25 college	Student rating (Purdue)	-.19	-.09	-.11	.21		
Sandiford (1936)	420 (approx.) student	Practice teaching grades	.14	(algebraic sum, /S /D -N -I)				
Ward & Kirk (1942)	95 student	Practice teaching rating (supervisor)	-.11	-.24	-.08	.04	.02	-.11
		Practice teaching rating (critic teacher)	.09	-.23	.01	.00	.08	-.11
Retan (1943)	152 with 2 to 5 yr. experience	Superintendent rating	Bernreuter & Pressey X.O 73 "stable" 75.3% "good & excellent" 79 "unstable" 91.9% "good & excellent."					
Gotham (1945)	57 elementary, 1- & 2-room	Pupil change	-.14	-.11	.04			
		Composite 13 teacher measures	.15	.19	.13			
		Composite (superintendent, supervisor, observer) 3 rating scales	.16	.01	.19			
		2 rating scales	.17	.04	.21			
Rolfe (1945)	47 elementary, 1- and 2-room	Pupil gain	-.14	-.11	.04			
Rostker (1945)	47 elementary, rural[a]	Pupil gain	-.28	-.23	.27	-.29	-.09	
Seagoe (1945)	31 student	Practice teaching rating	-.38	-.06				
Seagoe (1946)	25 elementary, 2 yr. experience	Administrator ranking (percentile)	-.30	-.26				
Espenschade (1948)	46 women, physical education, 1 yr. experience	Administrator rating	.09	-.23		.10		.28
			Bell Adjustment Inventory[b]					
Phillips (1935)	173 elementary & jr. high school	Average superintendent & principal ratings	-.08					
		Superintendent rating (sigma scores)	-.07					
Seagoe (1945)	31 student	Practice teaching rating	-.40					
Seagoe (1946)	25 elementary, 2 yr. experience	Administrator ranking (percentile)	-.31					
Jones (1946)	48 high school	Supervisor ratings	-.04					
	27 high school	Pupil gain	-.27					
Gould (1947)	92 with 1 yr. experience	Principal rating	.37[c]					
			Thurstone Personality Schedule					
Seagoe (1945)	31 student	Practice teaching rating	-.36					
Seagoe (1946)	25 student	Administrator ranking (percentile)	-.55					
Gould (1947)	113 with 1 yr. experience	Principal rating	.23[c]					
Ryans (1951)	275 elementary women	Composite rating of observers	Only 20% of items significant at .05 level or better					
			Rorschach					
Blair (1946)	205 in-service	None (experienced vs. inexperienced)	Multiple choice (No. of "poor" answers)	8.8% maladjusted (CR = 4.1, P = .01)				
	152 prospective		Multiple choice (No. of "poor" answers)	2.0% maladjusted				
Cooper & Lewis (1951)	30 highest & 30 lowest of 72 high school plus 153 student	Pupil rating	Absence of neurotic sign	.52 (tetrachoric)				
			Personality maladjustment	No relationship				
			Percentage of determinant	No significant difference				
			Human movement responses	Slightly higher for well-liked but not significant				

---

[a] Exclusive of 1- and 2-room schools.

[b] High scores on the Bell Adjustment Inventory indicate poor adjustment.

[c] Coefficient of contingency.

Table 35 (Cont.)

Investigator	Teacher sample	Measure of effectiveness	Correlation	
Callis (1952)	42 elementary	Student rating	Anxiety	.14
			Hostility	-.12
		Observer rating	Anxiety	.13
			Hostility	-.16
			Anxiety	-.04
			Hostility	-.05
			Minnesota Multiphasic	
Michaelis & Taylor (1951)	56 student women	Practice teaching rating (supervisor)	9 subscores	-.36 to .14
	31 student women	17 high-vs. 14 low-ranked teachers	Hysteria (only significant between significant score .01 & .05 of the 9 subtests)	
Callis (1952)	77 elementary	Student rating	Selected items:	.41
		Observer rating		.41
		Principal rating		.38
			Humm-Wadsworth Temperament Scale	
Grieder & Newburn (1942)	186 student	Practice teaching grades	Negative findings	
Seagoe (1945)	22 student	Practice teaching rating	Qualitative estimate	.63
	31 student	Practice teaching rating	"No-Count"	.30
Seagoe (1946)	25 with 2 yr. experience	Administrator ranking (percentile)	Qualitative estimate	.65
		Administrator ranking (percentile)	"No-Count"	.52
			Miscellaneous measures	
Brookover (1940)	39 high school	Student ratings (Purdue scale)	Student person-to-person interaction ratings	.64
Brookover (1945)	66 high school, male	Pupil gains in history information	Pupil rating - personal	Low signifi. neg. rel.
Cook & Leeds (1947)	100 "unselected"	Pupil rating - personal effectiveness	Leeds Teacher-Pupil Inventory	.45
Leeds (1950)		Principals' rating - personal effectiveness	Leeds Teacher-Pupil Inventory	.43
		Experts' rating - personal effectiveness	Leeds Teacher-Pupil Inventory	.49
Callis (1952)	77 elementary	Student rating	Minn. Teacher Attitude Inventory (MTAI)	.49
		Observer rating		.40
		Principal rating		.19

(d); for the "dominant" category at the .01 level for Factors (a), (d) for total rating, and for rating of pupil behavior (taken separately) for the "sociable" factor at the .05 level for Factors (a), (d), and rating of pupil behavior.

Other personality tests given to teachers have included the Pressey X-O Test (271), the Rudisill scale for measurement of the personality of elementary teachers (132), the Occupational Personality Inventory (101, 102), tests of Cattell's primary source traits (297), Cattell's 16 Personality Factor Test (189), Johnson Temperament Analysis, Minnesota Personality Scale, and Minnesota T-S-E Test (337). Correlation coefficients where reported tend to be low and are probably not significant except perhaps for some of those found by Schwartz (297). Using 34 teachers, he reports coefficients ranging from -.32 to .28 when tests of "primary source traits" were correlated with practice teaching rating, and coefficients from -.60 to .31 (*N* = 18) when the "primary source traits" were correlated with supervisors' ratings.

Lamke (189), in 1951, attempted to find out if the personalities of good and poor teachers as evaluated by Cattell's 16 Personality Factor Test were characteristically different. He used Fisher's discriminant function and factor analysis in the examination of his data. Results of the analysis by either method failed to reveal a characteristic personality pattern for either the good or the poor teachers. Lamke says the response patterns of the teachers studied on the 16 personality factor test suggest that "It is possible that personality traits need to be 'balanced in a certain way for the teacher to be superior. Lacking this balance', perhaps the teacher is likely to be only average; with a certain makeup she may be poor." Considering the results of the factor analysis of the responses to this test, Lamke concludes:

> "Using Cattell's terminology, it appears that good teachers are likely, more than poor teachers, to be gregarious, adventurous, frivolous, to have abundant emotional responses, strong artistic or sentimental interests, to be interested in the opposite sex, to be polished, fastidious and cool. Poor teachers are more likely than good teachers to be shy, cautious, conscientious, to lack emotional response and artistic or sentimental interests, to have a comparatively slight interest in the opposite sex, to be clumsy, easily pleased, and more attentive to people." (Lamke, Reference 189.)

Other measures, related to personality tests, which have been studied by a number of investigators, are those pertaining to various aspects of social adjustment. The results of the studies dealing with these variables are shown in Table 36. It will be seen that most of the correlation coefficients found between social adjustment measures and other measures of instructor effectiveness tend to cluster around zero. Some exceptions are evident in the case of the Washburne Social Adjustment Inventory and Jackson's Social Proficiency Test. Correlations ranged from .40 reported by Gotham to -.60 found by Schwartz when scores on the Washburne inventory were correlated with ratings. LaDuke obtained a correlation coefficient of -.37 when he correlated scores on the Jackson test with pupil gains. The extreme variability of results found with the Washburne Social Adjustment Inventory and the generally insignificant relationships shown by other "social" tests suggest that such measures have little to contribute as predictors of instructor effectiveness.

Results obtained with personality tests of teachers have in general shown wide variation when correlated with measures of teacher effectiveness. Correlations range from rather large positive or negative relationships to zero or near zero relationships depending upon the particular situation and the teacher measures used. There are many conceivable kinds of effectiveness even for teachers of the same subject or grade level in the same kind of community and therefore there will probably be different patterns of teacher personality for such effectiveness. As Lamke and others have pointed out, success in teaching may be a "balance" and to predict success it may be necessary to understand what is required for the balance. Study of the association of traits, one by one, with success will not suffice. The problem of determining the personality patterns of the effective teachers still remains unsolved, despite the fact that some so-called personality (and other) measures apparently show significant correlations (either positive or negative) with certain measures of instructor effectiveness. Carefully controlled, well-designed studies employing adequate numbers of instructors are needed to determine what measures or combinations of measures have definite predictive value. There is probably even a greater need for the development of adequate rationales, frameworks, and systems of hypotheses which are based on the best available theories concerning social interaction, interpersonal relationships, motivation, and learning. Through research effort these theories may then be related to specified dimensions of teacher personality and performance.

IMPLICATIONS FOR FURTHER RESEARCH

After scrutiny of several hundred research studies pertaining more or less directly to the identification of instructor effectiveness, the reviewers have arrived at certain conclusions with respect to the areas in

Table 36

Relation of Social Adjustment Measures to Measures of Instructor Effectiveness

Investigator	Teacher sample	Test used	Measure of effectiveness	Correlation
Morris (1929)	60 student	"Sympathy" (devised by author)	Practice teaching grade	.16
Barr, *et al*. (1935)	66 elementary	Moss Social Intell.	Pupil gain Superintendent rating (composite 7 scales)	.13 .10
Gotham (1945)	57 elementary, 1- and 2-room	Washburne Social Adjustment Inventory	Pupil gain Composite rating (3 scales) Composite rating (2 scales)	.06 .40 .40
Rolfe (1945)	47 elementary, 1- and 2-room	Washburne Social Adjustment Inventory	Pupil gain	.06
Rostker (1945)	28 elementary, rural[a]	Washburne Social Adjustment Inventory	Pupil gain	.14
Gould (1947)	113 with 1 yr. experience	Washburne Social Adjustment Inventory	Principal rating	.35[b]
Riesch (1949)	22 elementary	Washburne Social Adjustment Inventory	Pupil gain Superintendent rating	.02 -.00
Schwarts (1950)	18 with 2 yr. experience	Washburne Social Adjustment Inventory	Supervisor rating	-.60
	34 student	Washburne Social Adjustment Inventory	Practice teaching rating	-.2
LaDuke (1945)	31 elementary, 1-room	Jackson's Social Proficiency	Pupil gain (composite)	-.37
Seagoe (1945)	31 student	Willoughby Emotional Maturity Scale Strong Interest Blank M-F scale	Practice teaching rating Practice teaching rating	.19 .08
Seagoe (1946)	25 elementary, 2 yr. experience	Willoughby Strong M-F scale	Supervisor ranking (percentile)	.02 .08

---

[a] Exclusive of 1- and 2-room schools.

[b] Coefficient of contingency.

which further research is needed and in which the probabilities of securing worth-while results appear greatest. In certain other areas, however, the available studies seem to demonstrate beyond reasonable doubt that research has already proceeded for a considerable distance up a blind alley. The problems which in the opinion of the reviewers appear worthy of further research fall into both the main categories into which the review is organized those problems having to do with the search for more adequate criteria of instructor effectiveness and those problems concerning discovery or improvement of predictors of the criteria.

Criterion Research

The changes induced in the students by the instructor appear to constitute the most important component of any criteria of instructor effectiveness. As Orleans _et al_. (249), Evans (382), and others have pointed out, the ideal criterion of the effective instructor is probably a composite of several measures. For Air Force instructors it seems obvious that the relative gains in subject-matter knowledge of groups of students under different instructors should be a most important element in this composite. The Air Force technical schools because of the large numbers of personnel instructing in the same subject-matter fields offer an ideal situation in which to make a thorough investigation of this criterion.

The results obtained from any simple use of raw gains scores are certain to be misleading. The adequate use of the gains criterion requires the control of such variables as student aptitude, ability and motivation, the effects of distractions, diverse classroom conditions, cultural differences in different localities, and the like.

The reliability of a measure of instructor effectiveness should be the reliability of that effect on different or successive classes and not the split-half reliability determined from the same class in which situational and temporal variance (more properly reviewed as error variance) increases the estimated reliability. This involves rather elaborate design and statistical manipulation much beyond the scope of the average school system or the average supervisor's capabilities. As a practical measurement device, apart from its use in an experimental situation, the measurement of student gains affords a costly, unwieldy, and laborious method of evaluating instructors. If it can be shown that student gains correlated adequately with some other more easily obtained measures, these latter could be used for most research and administrative purposes as substitutes.

The demand continues for more objective measures to be used for instructor selection and evaluation. Precise methods of direct observation have been little used in determining instructor effectiveness, probably because of the inherent difficulties in their application. Such observations require study as potential predictors of other criteria of instructor performance; measures of observable behavior which turn out to be valid could then, in turn, be further used as criteria for future research or for practical application as evaluation indexes. Exploratory studies designed to investigate various techniques of instructor observation are thus urgently needed. The utilization of tape recorders, photographic, and other recording devices in connection with observation of instructors has not been thoroughly investigated. While some work has been directed toward observing instructors in a classroom situation there appear to be few, if any, studies of methods for making reliable observations of instructor and student behavior in the laboratory or shop. In this connection the methods of Olson and Wilkinson (248), by means of which they attempted to determine differences among teachers in terms of the amount and kind of verbal direction used in controlling behavior of elementary school pupils, appear worthy of further investigation. Their techniques, if modified to suit adult students, might well produce results of value in the evaluation of instructors and instructional methods in Air Force technical training schools. Observation to be of research value, however, must be repeatable by other scientists. Judgments of instructors that depend for their accuracy on the intuition or diagnostic skill of a lone observer are not adequate data for research. This may mean that every possibility of success is eliminated, but it still remains to be demonstrated that behaviors which can be reliably observed by different observers and which are reliably associated with different occasions (are typical of the instructor) are not related to effectiveness.

The relatively high coefficients obtained by Shannon (307), when he correlated student attention scores with scores on achievement tests, also suggest a lead which might prove useful if applied to students in an Air Force situation, despite Shannon's rather low opinion of his findings. (See the section on Objective Observation of Instructor Performance.)

There are, however, other aspects of the instructor's performance that may play some part in his over-all effectiveness as a member of a group with a common goal. For instance, the instructor has certain administrative and clerical responsibilities that, while they do not add to student gains, are important to the orderly administration of the training courses. Further, it is possible for instructors to contribute to a greater or lesser degree to improvement of the curriculum and to the development and promotion of better methods of presentation. Estimates of the extent to which different instructors make such contributions are probably best obtained from supervisors' ratings of instructors.

Additionally, it seems possible that the behaviors and expressed attitudes of the instructor could have a marked effect on the willingness of both his students and fellow instructors to work together to accomplish a group mission. In other words, the influence of the instructor on school morale may also be an aspect of his effectiveness. This aspect would probably best be reflected in ratings of the instructor made by his fellow instructors and by his students. Nothing is known of the amount of interrelationship or the extent of independent reliable variance likely to be found in such measures in Air Force schools. Considerable research effort would be necessary to determine the weightings that should be used in any composite criterion of instructor effectiveness.

The general unsatisfactory nature of past rating methods has stimulated the search for more satisfactory techniques. Among rating methods the forced-choice technique evidently offers some promise for operational use since it tends to reduce biasability. Considerable research would be required, however, to determine the value of forced-choice scales devised for use by student raters, fellow teachers, or as self-rating scales. It must be determined also whether or not repeated ratings on forced-choice forms, like those on graphic scales, tend to become progressively more lenient and less valid.

Little practical use has been made of fellow teacher ratings in civilian institutions. While an instructor's opinions of his fellow instructors may be biassed, it is more than probable that through his day-to-day close contacts with them he knows what kind of instructors they are. His relationships with his fellow instructors being different from those of the supervisors or students will enable him to know them in a somewhat different way and his judgments of them will be based on this different point of view. Peer ratings of instructors in the Air Force should receive further investigation, either through forced-choice or other methods, in the expectation that they might be used to corroborate supervisor ratings or as a part of a composite to bring about a more adequate rating of instructors than supervisor ratings used alone.

Student ratings are being used more and more widely in civilian schools and colleges, a trend in keeping with the present day tendencies to give greater emphasis to the democratic process in education. The argument is frequently advanced that in the Air Force technical schools the phases are so short that the student has insufficient time to get well enough acquainted with his instructor to make adequate judgment of him. The total hours an Air Force technical school student spends with his instructor, however, are often considerably greater than the time a college student spends with his instructor during a one semester college course. As in the case of peer ratings, student ratings have played no great part in the evaluation of Air Force technical training school instructors. Thorough study would be required to determine their utility for self-improvement of instructors and also to discover their value as a criterion per se or as a predictor of gains or other criteria of instructor effectiveness.

It is possible that the use of a composite criterion will obscure patterns and significant elements or specific aspects of effectiveness. It may be difficult to add together, say by means of regression equation techniques, different components of teacher effectiveness so that a high degree of one component is allowed to counterbalance a low degree of another, when both may be equally important in their own way. Thus, it may be necessary to develop new ways of combining or otherwise utilizing several criteria. The development of such a composite will require the best available judgment on the part of psychologists and school administrators as to the relative weights to be assigned considering the interrelations found.

Since many of the studies reviewed have been concerned with ratings, in the foregoing discussion of implications for further research on criteria, the reviewers have emphasized methodological considerations. The major problems of research on criteria, however, may not be methodological but rather conceptual or definitional problems. The objectives of training programs need to be defined, students' achievements of these objectives insofar as they can be measured need to be ascertained, and the effects of instructors on these achievements need to be isolated.

It is contended by some educational authorities that no kind of rating on any kind of scale by any kind of person is likely to provide an acceptable criterion until it can be shown to be related to student change in the direction of the educational objectives of the school or training program. This is an extreme position which would appear to rule out, as unacceptable, ratings which tend to show negligible correlations with student gains. While it appears reasonable that measurable student changes should constitute a part (perhaps the largest part) of a total criterion of teacher effectiveness, it is also possible that ratings may reflect areas of effectiveness not directly measurable. The question of whether ratings are acceptable as a part of a total criterion depends on whether there are logical grounds for believing that the teacher can contribute to the accomplishment of school objectives in addition to his effects on his own students. If it seems possible for teachers to contribute differentially to the group efforts through work on the curriculum, through development of improved methods, through their influence on group morale, etc., then these contributions should be a part of any total criterion of effectiveness. If it likewise seems possible that ratings might reflect the quality of a teacher's participation in the group effort, then the use of ratings as an element in a total criterion of effectiveness is justified.

To the reviewers the major problem connected with ratings is not the justification of their use, but rather the improvement of their accuracy.

Predictor Research

Research on predictors necessitates formulation of hypotheses and the development of conceptual frameworks based on the best available psychological and educational theories. These hypotheses will reflect the rationale that certain traits or behavior of an instructor may be expected to be related to and hence may be used as predictors of instructor competence. For example, hypotheses might be set up with respect to the relation of instructors' intelligence to instructor effectiveness for different kinds of subject matter. Similar hypotheses might be generated for age, experience, extracurricular activities, sex, verbal facility, and other instructor variables.

The differential relations of instructor intelligence to instructor effectiveness for different kinds of subject matter should be determined. Likewise the optimal relations between instructor intelligence and the aptitude and experience levels of students should be investigated. The student gains criterion might be used to determine the value of intelligence as a predictor of instructors' competence in courses of differing complexity. It is quite possible that the intelligence factor when used with other instructor measures might contribute materially to an instructor selection battery.

A number of investigators have shown a relationship between instructor effectiveness and age or experience which appears to be curvilinear. Teachers tend to reach maximum rated efficiency after five or more years of teaching experience. In the Air Force, however, extremely few (approximately two per cent) airman instructors remain in a teaching assignment for as long as five years. If the Air Force is indeed losing the majority of airman instructors before they reach their period of maximum efficiency, a change in policy might be anticipated.

The evidence suggests that the kind and number of activities a teacher has engaged in may have some relation to his effectiveness as a teacher. This finding, as shown with respect to certain specific extracurricular activities in the case of some civilian school teachers, might also apply to Air Force technical school instructors. A study might be made to determine if past participation and interest in specific activities (or in many varied activities) are related to an instructor's success in training student airmen to become proficient in varied technical school specialties.

No fundamental differences in instructional effectiveness between men and women teachers have been demonstrated. Although these findings were obtained in quite different training situations from Air Force technical courses, the possibility of utilizing WAF instructors should not be overlooked.

The rather interesting findings of McCoard (210), with respect to verbal facility suggest several potentially fruitful areas of research: (a) to determine the relationship between the verbal facility and technical information an instructor shows in the classroom as compared with his ability to demonstrate equipment and procedures in the technical laboratory or shop; (b) to determine the extent to which an instructor's ability to organize and present verbal material is related to the subject-matter gains of his students; (c) to find out if verbal facility can be measured and used as part of the instructor selection procedure.

The investigations of factor analysis of instructor abilities so far available are somewhat vitiated due to inadequacies of criteria, measuring devices, or numbers of cases used. A more adequately designed investigation might yield factorial results which might prove of considerable value toward the solution of instructor selection and evaluation problems in the Air Force.

The personality patterns of the successful instructors have not yet been determined. This does not mean, however, that this approach should be abandoned. Carefully controlled, well-designed experiments employing adequate numbers of instructors would be needed in which plausible measures or combinations of such measures are investigated. Certain tests used in preliminary studies have shown promise. These should be used in more thoroughgoing experiments. The search should continue also for new and untried measuring instruments in the hope that some device will be discovered which will enable the Air Force to predict teaching success of instructors in training and to evaluate instructors on the job.

It should be pointed out that the importance of many of the problems suggested by this Research Bulletin has been recognized by the Air Force Personnel and Training Research Center, and preliminary experiments in several of these areas are now underway.

BIBLIOGRAPHY

1. ALBERT, H.R. An analysis of teacher rating by pupils in San Antonio, Texas. *Educ. Adm. Supervis.*, 1941, 27, 267-274.

2. ALLEN, H.F. Earmarks of a good teacher. *Amer. Sch. Bd J.*, 1938, 96 (3), 25-26; 92.

3. ALMY, H.C., and SORENSON, H. A teacher-rating scale of determined reliability and validity. *Educ. Adm. Supervis.*, 1930, 16, 179-186.

4. AMATORA, S.M. A diagnostic teacher-rating scale. *J. Psychol.*, 1950, 30, 395-399.

5. ANDERSEN, W.N. The selection of teachers. *Educ. Adm. Supervis.*, 1917, 3, 83-90.

6. ANDERSON, H.H., and BREWER, HELEN M. Studies of teachers' classroom personalities, I: Dominative and socially integrative behavior of kindergarten teachers. *Appl. Psychol. Monogr.*, 1945, No. 6.

7. ANDERSON, H.H., and BREWER, J.E. Studies of teachers' classroom personalities, II: Effects of teachers' dominative and integrative contacts on children's classroom behavior. *Appl. Psychol. Monogr.*, 1946, No. 8.

8. ANDERSON, H.H., BREWER, J.E., and REED, MARY FRANCES. Studies of teachers' classroom personalities, III: Follow-up studies of the effects of dominative and integrative contacts on children's behavior. *Appl. Psychol. Monogr.*, 1946, No. 11.

9. ANDERSON, H.J. Correlation between academic achievement and teaching success. *Elem. Sch. J.*, 1931, 32, 22-29.

10. ARMENTROUT, W.D. The rating of teachers by training teachers and superintendents. *Elem. Sch. J.*, 1928, 28, 511-516.

11. ASCH, S.E. Forming impressions of personality. *J. abnorm. soc. Psychol.*, 1946, 41, 258-290.

12. BACH, J.O. Practice teaching success in relation to other measures of teaching ability. *J. exp. Educ.*, 1952, 21, 57-80.

13. BAIER, D.E. Reply to Travers' A critical review of the validity and rationale of the forced-choice technique. *Psychol. Bull.*, 1951, 48, 421-434.

14. BAIRD, J., and BATES, G. The basis of teacher rating. *Educ. Adm. Supervis.*, 1929, 15, 175-183.

15. BARKER, M. ELIZABETH. Summary of the relation of personality adjustments of teachers to their efficiency in teaching. *J. educ. Res.*, 1948, 41, 664-675.

16. BARR, A.S. *Characteristic differences in the teaching performance of good and poor teachers of the social studies*. Bloomington, Ill: Public School Publishing Co., 1929.

17. BARR, A.S. Teaching competencies. In W.S. Monroe (Ed.), *Encyclopedia of educational research*. New York: Macmillan, 1950. Pp. 1446-1454.

18. BARR, A.S., BECHDOLT, B.V., COXE, W.W., GAGE, N.L., ORLEANS, J.S., REMMERS, H.H., and RYANS, D.G. Report of the committee on the criteria of teacher effectiveness. *Rev. Educ. Res.*, 1952, 22, 238-263.

19. BARR, A.S., and EMANS, L.M. What qualities are prerequisite to success in teaching? *Nation's Sch.*, 1930, 6 (3), 60-64.

20. BARR, A.S., TORGERSON, T.L., JOHNSON, C.E., LYON, V.E., and WALVOORD A.C. The validity of certain instruments employed in the measurement of teaching ability. In Helen M. Walker (Ed.), *The measurement of teaching efficiency*. New York: Macmillan, 1935. Pp. 73-141.

21. BARTHELMESS, HARRIET M., and BOYER, P.A. A study of the relation between teaching efficiency and amount of college credit earned while in service. *Educ. Adm. Supervis.*, 1928, 14, 521-535.

22. BARTHELMESS, HARRIET M., and BOYER, P.A. Relation between teaching and amount of college credit earned while in service. *Penn. Sch. J.*, 1929, 77, 291.

23. BATHURST, J.E. Do teachers improve with experience? *Personnel J.*, 1928, 7, 54-57.

24. BATHURST, J.E. Relation of efficiency to experience and age among elementary teachers. *J. educ. Res.*, 1929, 19, 314-316.

25. BAXTER, BERNICE. *Teacher-pupil relationships*. New York: Macmillan, 1941.

26. BEAUMONT, H. A suggested method for measuring the effectiveness of teaching introductory courses. *J. educ. Psychol.*, 1938, 29, 607-612.

27. BELL, S. A study of the teacher's influence. *Pedag. Semin.*, 1900, 7, 492-525.

28. BELL, VIOLA M. Traits of teachers. *Educ. Res. Bull.*, 1932, 11, 281-286.

29. BENDIG, A.W. Inter-judge vs intra-judge reliability in the order-of-merit method. Amer. J. Psychol., 1952, 65, 84-89.

30. BENDIG, A.W. A preliminary study of the effect of academic level, sex, and course variables on student rating of psychology instructors. J. Psychol., 1952, 34, 21-26.

31. BENT, R.K. Relationships between qualifying examinations, various other factors, and student teaching performance at the University of Minnesota. J. exp. Educ., 1937, 5, 251-255.

32. BETTS, G.L. The education of teachers evaluated through measurement of teaching ability. In National survey of the education of teachers. U.S. Off. Educ. Bull., 1933, No. 10 (5). Pp. 87-153.

33. BETTS, G.L. Pupil achievement and the NS trait in teachers. In Helen M. Walker (Ed.), The measurement of teaching efficiency. New York: Macmillan, 1935. Pp. 144-237.

34. BIMSON, O.H. Do "good" teachers produce "good" results? North Central Ass. Quart., 1937, 12, 271-276.

35. BIRD, GRACE E. Pupils' estimates of teachers. J. educ. Psychol., 1917, 8, 35-40.

36. BIRKELO, C.P. What characteristics in teachers impress themselves most upon elementary and high school students? Educ. Adm. Supervis., 1929, 15, 453-456.

37. BLAIR, G.M. Personality adjustments of teachers as measured by the multiple choice Rorschach test. J. educ. Res., 1946, 39, 652-657.

38. BLISS, W.B. How much mental ability does a teacher need? J. educ. Res., 1922, 6, 33-41.

39. BOARDMAN, C.W. Professional tests as measures of teaching efficiency in high school. Teach. Coll. Contr. Educ., 1928, No. 327.

40. BOARDMAN, C.W. An analysis of pupil ratings of high school teachers. Educ. Adm. Supervis., 1930, 16, 440-446.

41. BOLTON, F.B. Evaluating teaching effectiveness through the use of scores on achievement tests. J. educ. Res., 1945, 38, 691-696.

42. BOND, J.A. Strengths and weaknesses of student teachers. J. educ. Res., 1951, 45, 11-22.

43. BOOK, W.F. The high school teacher from the pupil's point of view. Pedag. Semin., 1905, 12, 239-288.

44. BOSSING, N.L. Teacher-aptitude tests and teacher selection. U. S. Off. Educ. Bull., 1931, No. 12, 117-133.

45. BOUSFIELD, W.A. Students' ratings of qualities considered desirable in college professors. Sch. & Soc., 1940, 51, 253-256.

46. BOWDEN, A.O. Change--the test of teaching. Sch . & Soc., 1934, 40, 133-136.

47. BOWMAN, E.C. Pupil ratings of student teachers. Univer. Ind. Educ. Sch. Bull., 1934, 11, 28-37. Also Educ. Adm. Supervis., 1934, 20, 141-146.

48. BOYCE, A.C. Qualities of merit in secondary school teachers. J. educ. Psychol., 1912, 3, 144-157.

49. BOYCE, A.C. Methods for measuring teachers' efficiency. Nat. Soc. Stud. Educ. Yearb., 1915, 14 (Part II), 1-83.

50. BRADLEY, GLADYCE H. What do college students like and dislike about college teachers and their teaching? Educ. Adm. Supervis., 1950, 36, 113-120.

51. BRANDT, W.J. A follow-up of some earlier Wisconsin studies of teaching ability. J. exp. Educ., 1949, 18, 1-29.

52. BRECKINRIDGE, ELIZABETH. A study of the relation of preparatory school records and intelligence test scores to teaching success. Educ. Adm. Supervis., 1931, 17, 649-660.

53. BREED, F.S. Factors contributing to success in college training. J. educ. Res., 1927, 16, 247-253.

54. BROOKOVER, W.B. Person-person interaction between teachers and pupils and teaching effectiveness. J. educ. Res., 1940, 34, 272-287.

55. BROOKOVER, W.B. The relation of social factors to teaching ability. J. exp. Educ., 1945, 13, 191-205.

56. BROOM, M.E. The predictive value of three specified factors for success in practice-teaching. Educ. Adm. Supervis., 1929, 15, 25-29.

57. BROOM, M.E. A note on predicting teaching success. Educ. Adm. Supervis., 1932, 18, 64-67.

58. BROOM, M.E., and AULT, J.W. How may we measure teaching success? Educ. Adm. Supervis., 1932, 18, 250-256.

59. BROWNELL, W.A. A critique of research on learning and on instruction in the school. Nat. Soc. Stud. Educ., 1951, 50 (1), 52-66.

60. BRUBACHER, A.R. Why teachers fail. New York State Educ., 1931, 18, 593-600.

61. BRYAN, R.C. Pupil rating of secondary school teachers. Teach. Coll. Contr. Educ., 1937, No. 708.

62. BRYAN, R.C. Pupil ratings of secondary-school teachers. Sch. Rev., 1938, 46, 357-367.

63. BRYAN, R.C. Eighty-six teachers try evaluating student reactions to themselves. Educ. Adm. Supervis., 1941, 27, 513-526.

64. BRYAN, R.C. Reliability, validity, and needfulness of written student reactions to teachers. Educ. Adm. Supervis., 1941, 27, 655-665.

65. BRYAN, R.C. Why student reactions to teachers should be evaluated. Educ. Adm. Supervis., 1941, 27, 590-603.

66. BRYAN, R.C. Benefits reported by teachers who obtained written student reactions. Educ. Adm. Supervis., 1942, 28, 69-75.

67. BUCKINGHAM, B.R. Opinion and practice as to the rating of teachers. Educ. Res. Bull., 1922, 1, 171-174.

68. BUELLESFIELD, H. Causes of failures among teachers. Educ. Adm. Supervis., 1915, 1, 439-452.

69. BUTLER, F.A. Prediction of success in practice teaching. Educ. Adm. Supervis., 1935, 21, 448-456.

70. BUTSCH, R.L.C. The two-factor theory applied to measurement of teacher traits. Educ. Adm. Supervis., 1932, 18, 257-276.

71. CALLIS, R., et al. Studies on the effectiveness of teaching. In Differential characteristics of the more effective and less effective teachers. A summary report of nine studies. Washington: Division of Field Studies and Training, Extension Service, U. S. Department of Agriculture, February 1953.

72. CATTELL, R.B. The assessment of teaching ability. Brit. J. educ. Psychol., 1931, 1, 48-72.

73. CHAMPLIN, C.D. The preferred college professor. Sch. & Soc., 1928, 27, 175-177.

74. CHARTERS, W.W., and WAPLES, D. The commonwealth teacher-training study. Chicago: Univer. Chicago Press, 1929.

75. CHEYDLEUR, F.D. Judging teachers of basic French courses by objective means at the University of Wisconsin--1919-1943. J. educ. Res., 1945, 39, 161-192.

76. CLAPP, F.L. Scholarship in relation to teaching efficiency. Sch. Rev. Monogr., 1915, 6, 64-70.

77. CLEM, O.M. What do my students think about my teaching? Sch. & Soc., 1930, 31, 96-100.

78. CLINTON, R.J. Qualities college students desire in college instructors. Sch. & Soc., 1930, 32, 702.

79. COLLINS, E.A. Relation of intelligence to success in teaching. Sch. & Community, 1930, 16, 155-157.

80. COOK, W.W., and LEEDS, C.H. Measuring the teaching personality. Educ. psychol. Measmt, 1947, 7, 399-410.

81. COOKE, D.H. How do teachers rate themselves? Educ. Adm. Supervis., 1937, 23, 473-476.

82. COOPER, HAZEL E. Correlation of the grades in practice teaching received by seniors in a college for teachers, with their scores in the Thurstone Group Intelligence Test. Pedag. Semin., 1924, 31, 176-182.

83. COOPER, J.G., and LEWIS, R.B. Quantitative Rorschach factors in the evaluation of teacher effectiveness. J. educ. Res., 1951, 44, 703-707.

84. COPPER, F.LeR. Who is a good teacher? Education, 1928, 49, 111-117.

85. COURTIS, S.A. Standards of teaching ability. Educ. Rev., 1921, 62, 183-186.

86. COURTIS, S.A. The measurement of the effect of teaching. Sch. & Soc., 1928, 28, 52-56; 84-88.

87. COXE, W.W., and CORNELL, ETHEL L. The prognosis of teaching ability of students in New York state normal schools. Univer. State of New York Bull., 1933, No. 1033.

88. COY, GENEVIEVE L. A study of various factors which influence the use of the accomplishment quotient as a measure of teaching efficiency. J. educ. Res., 1930, 21, 29-42.

89. CRABBS, LELAH M. Measuring efficiency in supervision and teaching. Teach. Coll. Contr. Educ., 1925, No. 175.

90. DALDY, DOROTHY M. A study of adaptability in a group of teachers. *Brit. J. educ. Psychol.*, 1937, 7, 1-22.

91. DANIEL, J.McT. *Excellent teachers. Their qualities and qualifications. Report of the investigation of educational qualifications of teachers in South Carolina.* Columbia: Univer. of South Carolina, 1944.

92. DAVENPORT, K. An investigation into pupil rating of certain teaching practices. *Purdue Univer. Stud. higher Educ.*, 1944, No. 49.

93. DAVIES, J.E. What are the traits of the good teacher from the standpoint of junior high school pupils? *Sch. & Soc.*, 1933, 38, 649-652.

94. DAVIS, C.O. The high school as judged by its students. *Proc. North Central Ass. Coll. Sec. Schs.*, 1924, 29, 71-144.

95. DAVIS, C.O. Our best teachers. *Sch. Rev.*, 1926, 34, 754-759. Also *Sch. & Soc.*, 1926, 24, 240-243.

96. DAVIS, H.McV. The use of state high school examinations as an instrument for judging the work of teachers. *Teach. Coll. Contr. Educ.*, 1934, No. 611.

97. DAVIS, S.B., and FRENCH, L.C. Teacher rating. *Pittsburgh Univer. Sch. Educ. J.*, 1928, 3, 57; 60-64.

98. DAY, L.C. The teaching quotient. *Elem. Sch. J.*, 1933, 33, 604-607.

99. DEAN, C.D. Current trends in rating student-teachers. *Educ. Adm. Supervis.*, 1939, 25, 687-694.

100. DODD, M.R. A study of teaching aptitude. *J. educ. Res.*, 1933, 26, 517-521.

101. DODGE, A.F. What are the personality traits of the successful teacher? *J. appl. Psychol.*, 1943, 27, 325-337.

102. DODGE, A.F. A study of the personality traits of successful teachers. *Occupations*, 1948, 27, 107-112.

103. DOLCH, E.W., Jr. Pupils' judgments of their teachers. *Pedag. Semin.*, 1920, 27, 195-199.

104. DOMAS, S.J. *Report of an exploratory study of teacher competence.* Cambridge, Mass.: Peabody House, 1950.

105. DRUCKER, A.J., and REMMERS, H.H. Do alumni and students differ in their attitudes toward instructors? *J. educ. Psychol.*, 1951, 42, 129-143.

106. ELLIOTT, E.C. Outline of a tentative scheme for the measurement of teaching efficiency. Madison, Wis.: Democrat Printing Co., 1910.

107. ELLIOTT, E.C. How shall the merit of teachers be tested and recorded? Educ. Adm. Supervis., 1915, 1, 291-299.

108. ENGELHART, M.D., and TUCKER, L.R. Traits related to good and poor teaching. Sch. Rev., 1936, 44, 28-33.

109. ESPENSCHADE, A. Selection of women major students in physical education. Res. Quart. Amer. Ass. Hlth., 1948, 19, 70-76.

110. FERGUSON, H., and HOVDE, H.O. Improving teaching personality by pupil rating. Sch. Rev., 1942, 50, 439-443.

111. FICHANDLER, A. A study in self-appraisal. Sch. & Soc., 1916, 4, 1000-1002.

112. FLANAGAN, J.C. A preliminary study of the validity of the 1940 edition of the National Teacher Examinations. Sch. & Soc., 1941, 54, 59-64.

113. FLANAGAN, J.C. Critical requirements: A new approach to employee evaluation. Personnel Psychol., 1949, 2, 419-426.

114. FLANAGAN, J.C. The critical requirements approach to educational objectives. Sch. & Soc., 1950, 71, 321-324.

115. FLESHER, W.R. Inferential study rating of instructors. Educ. Res. Bull., 1952, 31, 57-62.

116. FLINN, VEE. Teacher rating by pupils. Educ. Method, 1932, 11, 290-294.

117. FLORY, C.D. Personality rating of prospective teachers. Educ. Adm. Supervis., 1930, 16, 135-143.

118. FORDYCE, C. Note on the correlations between general teaching power and some specific teaching qualities. Nat. Soc. Stud. Educ. Yearb., 1919, 18 (Part I), 349-351.

119. FRASIER, C.M. Intelligence as a factor in determining student teaching success. Educ. Adm. Supervis., 1929, 15, 623-629.

120. FRASIER, G.W. Teaching as rated by pupils, school officers and other teachers. Hawaii educ. Rev., 1927, 15, 170-171; 176.

121. FREDERICK, R.W., and HOLLISTER, F.C. The relationship between the academic success of pupils and the practice teaching grade received by their teachers. Educ. Adm. Supervis., 1934, 20, 468-471.

122. FREYD, M. The graphic rating scale. J. educ. Psychol., 1923, 14, 83-102.

123. FRITZ, M.F. The variability of judgment in the rating of teachers by students. Educ. Adm. Supervis., 1926, 12, 630-634.

124. FRY, J.C. All superior officers. Infantry J., 1948, 63, 21-26.

125. FULLER, ELIZABETH M. The use of measures of ability and general adjustment in the preservice selection of nursery school-kindergarten-primary teachers. J. educ. Psychol., 1946, 37, 321-334.

126. GALT, W.E., and GRIER, D.J. Evaluation and selection of flying instructors. In N. E. Miller (Ed.), Psychological research on pilot training. AAF Aviation Psychology Program Research Report No. 8, 1947, pp. 289-351.

127. GEORGES, J.S. Determining instructional efficiency. Sch. Rev., 1931, 39, 64-66.

128. GIESE, W.J., and STEVENS, S.N. The application blank can be made predictive of teaching success. Amer. Sch. Bd J., 1939, 98 (4), 23-25.

129. GOODENOUGH, FLORENCE L. Mental testing, its history, principles, and applications. New York: Rinehart, 1949.

130. GOODENOUGH, FLORENCE L., FULLER, ELIZABETH M., and OLSON, EDNA. The use of the Goodenough Speed-of-Association Test in the preservice selection of nursery school-kindergarten-primary teachers. J. educ. Psychol., 1946, 37, 335-346.

131. GOODHARTZ, A.S. Student attitudes and opinions relating to teaching at Brooklyn College. Sch. & Soc., 1948, 68, 345-349.

132. GOTHAM, R.E. Personality and teaching efficiency. J. exp. Educ., 1945, 14, 157-165.

133. GOULD, G. The predictive value of certain selective measures. Educ. Adm. Supervis., 1947, 33, 208-212.

134. GRAVES, NELLIE H. What traits do high school pupils admire in teachers? High Sch. J., 1925, 8, 103-105.

135. GREENE, H.W. A comparison of student ratings, administrative ratings, ratings by colleagues, and relative salaries as criteria of teaching excellence. W. Va. State Coll. Bull., 1933, No. 5.

136. GRIEDER, C., and NEWBURN, H.K. Temperament in prospective teachers. J. educ. Res., 1942, 35, 683-693.

137. GRIM, P.R., and HOYT, C.J. Appraisal of teaching competency. Educ. Res. Bull., 1952, 31, 85-91.

138. GUILFORD, J.P. Psychometric methods. New York: McGraw-Hill, 1936.

139. (Omitted)

140. GUTHRIE, E.R. Measuring student opinion of teachers. Sch. & Soc., 1927, 25, 175-176.

141. HAGGARD, W.W. Some freshmen describe the desirable college teacher. Sch. & Soc., 1943, 58, 238-240.

142. HAMPTON, NELLIE DELIGHT. An analysis of supervisory ratings of elementary teachers graduated from Iowa State Teachers College. J. exp. Educ., 1951, 20, 179-215.

143. HAMRIN, S.A. A comparative study of ratings of teachers-in-training and teachers-in-service. Elem. Sch. J., 1927, 28, 39-44.

144. HARDESTY, C.D. Can teaching success be rated? Nation's Sch., 1935, 15 (1), 27-28.

145. HART, F.W. Teachers and teaching by ten thousand high-school seniors. New York: Macmillan, 1936.

146. HARTMANN, G.W. Measuring teaching efficiency among college instructors. Arch. Psychol., 1933, 23, No. 154.

147. HATCHER, MATTIE L. Qualities of personality compared with success in practice-teaching. Peabody J. Educ., 1934, 11, 246-253.

148. HEILMAN, J.D., and ARMENTROUT, W.D. The rating of college teachers on ten traits by their students. J. educ. Psychol., 1936, 27, 197-216.

149. HELLFRITZSCH, A.G. A factor analysis of teacher abilities. J. exp. Educ., 1945, 14, 166-199.

150. HENRIKSON, E.H. Comparisons of ratings of voice and teaching ability. J. educ. Psychol., 1943, 34, 121-123.

151. HENRIKSON, E.H. Some relations between personality, speech characteristics and teaching effectiveness of college teachers. Speech Monogr., 1949, 16, 221-226.

152. HERDA, F.J. Some aspects of the relative instructional efficiency of men and women teachers. J. educ. Res., 1935, 29, 196-203.

153. HIGGINS, SISTER. Reducing the variability of supervisors' judgments: An experimental study. *John Hopkins Univer. Stud. Educ.*, 1936, No. 23.

154. HIGHLAND, R.W., and BERKSHIRE, J.R. *A methodological study of forced-choice performance rating.* San Antonio, Tex.: Human Resources Research Center, Lackland Air Force Base, May 1951. (*Research Bulletin* 51-9.)

155. HILL, C.W. The efficiency ratings of teachers. *Elem. Sch. J.*, 1921, 21, 438-443.

156. HILL, L.B. Teaching qualities in former graduates as guides in improving student-teaching. *Educ. Adm. Supervis.*, 1929, 15, 362-366.

157. HINES, H.C. Merit systems in the larger cities. *Amer. Sch. Bd J.*, 1924, 68 (6), 52; 111-112; 115.

158. HOPPOCK, R. N. Y. U. students grade their professors. *Sch. & Soc.*, 1947, 66, 70-72.

159. HUCKLEBERRY, A.W. The relationship between change in speech proficiency and change in student teaching proficiency. *Speech Monogr.*, 1950, 17, 378-389.

160. HUDELSON, E. The validity of student rating of instructors. *Sch. & Soc.*, 1951, 73, 265-266.

161. HULT, ESTHER. Study of achievement in educational psychology. *J. exp. Educ.*, 1945, 13, 174-190.

162. IRWIN, CLAIRE C., and IRWIN, J.R. An appraisal of certain teacher traits by representative high school students. *Cath. Educ. Rev.*, 1949, 47, 667-675.

163. JACOBS, C.L. The relation of the teacher's education to her effectiveness. *Teach. Coll. Contr. Educ.*, 1928, No. 277.

164. JAMES, H.W. Causes of teacher-failure in Alabama. *Peabody J. Educ.*, 1930, 7, 269-271.

165. JARECKE, W.H. Evaluating teaching success through the use of teaching judgment test. *J. educ. Res.*, 1952, 45, 683-694.

166. JAYNE, C.D. A study of the relationship between teaching procedures and educational outcomes. *J. exp. Educ.*, 1945, 14, 101-134.

167. JENSEN, A.C. Determining critical requirements for teachers. *J. exp. Educ.*, 1951, 20, 79-85.

168. JERSILD, A.T. Characteristics of teachers who are "liked best" and "disliked most." J. exp. Educ., 1940, 9, 139-151.

169. JOHNS, W.B., and WORCESTER, D.A. The value of the photograph in the selection of teachers. J. appl. Psychol., 1930, 14, 54-62.

170. JOHNSTON, J.H. Teacher rating in large cities. Sch. Rev., 1916, 24, 641-647.

171. JONES, E.S. The prediction of teaching success for the college student. Sch. & Soc., 1923, 18, 685-690.

172. JONES, R.DeV. The prediction of teaching efficiency from objective measures. J. exp. Educ., 1946, 15, 85-99.

173. JORDAN, F. A study of personal and social traits in relation to high-school teaching. J. educ. Sociol., 1929, 3, 27-43.

174. KELLY, R.L., and ANDERSON, RUTH E. Great teachers and some methods of producing them. J. educ. Res., 1929, 20, 22-30.

175. KEPHART, A.P. What kind of a teacher? Amer. Sch. Bd J., 1922, 64 (3), 47-48; 84; 87.

176. KING, LeR.A. The present status of teacher rating. Amer. Sch. Bd J., 1925, 70 (2), 44-46; 154; 157.

177. KLOPP, W.J. Evaluation of teacher traits by vacation-school pupils. Sch. Rev., 1929, 37, 457-459.

178. KNIGHT, F.B. Qualities related to success in teaching. Teach. Coll. Contr. Educ., 1922, No. 120.

179. KNIGHT, F.B. and FRANZEN, R.H. Pitfalls in rating schemes. J. educ. Psychol., 1922, 13, 204-213.

180. KNUDSEN, C.W., and STEPHENS, STELLA. An analysis of fifty-seven devices for rating teaching. Peabody J. Educ., 1931, 9, 15-24.

181. KRATZ, H.E. Characteristics of the best teacher as recognized by children. Pedag. Semin., 1896, 3, 413-418.

182. KRINER, H.L. Pre-training factors predictive of teacher success. Penn. State Coll. Stud. Educ., 1931, No. 1.

183. KRINER, H.L. Preliminary report on a five-year study of teachers college admissions. Educ. Adm. Supervis., 1933, 19, 691-695.

184. KRINER, H.L. Second report on a five-year study of teachers college admissions. Educ. Adm. Supervis., 1935, 21, 56-60.

185. KRINER, H.L. Five-year study of teacher college admissions. Educ. Adm. Supervis., 1937, 23, 192-199.

186. KROUS, G.T. A study of traits and qualities of teachers and their effectiveness in teaching, based upon the estimates of their students. In Stanford Univer., Abstracts of dissertations . . . 1934-35. Pp. 182-185.

187. KYTE, G.C. The problem teacher in the grades--a composite picture. Nation's Sch., 1932, 9 (5), 55-60.

188. LaDUKE, C.V. The measurement of teaching ability. Study number three. J. exp. Educ., 1945, 14, 75-100.

189. LAMKE, T.A. Personality and teaching success. J. exp. Educ., 1951, 20, 217-259.

190. LAMSON, EDNA E. Some college students describe the desirable college teacher. Sch. & Soc., 1942, 56, 615.

191. LANCELOT, W.H. Standards of measurement of teaching ability. Sch. & Soc., 1931, 34, 236-238.

192. LANCELOT, W.H. A study of teaching efficiency as indicated by certain permanent outcomes. In Helen M. Walker (Ed.), The measurement of teaching efficiency. New York: Macmillan, 1935. Pp. 3-69.

193. LANG, A.R. A study of amount of training and experience as objective factors entering into a basis for determining teachers' salaries. Unpublished doctor's dissertation, Stanford Univer., 1924.

194. LARSON, A.H., and MARZOLF, S.S. Attitude of teachers college students toward teaching. Educ. Adm. Supervis., 1943, 29, 434-438.

195. LaRUE, D.W. Emotional differences between superior and inferior teachers. Nat. elem. Principal, 1936, 15, 395-401.

196. LAWTON, J.A. A study of factors useful in choosing candidates for the teaching profession. Brit. J. educ. Psychol., 1939, 9, 131-143.

197. LAYCOCK, S.R. The Bernreuter Personality Inventory in the selection of teachers. Educ. Adm. Supervis., 1934, 20, 59-63.

198. LEEDS, C.H. A scale for measuring teacher-pupil attitudes and teacher-pupil rapport. Psychol. Monogr., 1950, 64, No. 6 (Whole No. 312).

199. LEIPOLD, L.E. Teacher traits that pupils like or dislike. *Clearing House*, 1949, 24, 164-166.

200. LIGHT, U.L. High-school pupils rate teachers. *Sch. Rev.*, 1930, 38, 28-32.

201. LINDGREN, H.C. The incomplete sentences tests as a means of course evaluation. *Educ. psychol. Measmt*, 1952, 12, 217-225.

202. LINDQUIST, E.F. *Educational measurement*. Washington: American Council on Education, 1951.

203. LINS, L.J. The prediction of teaching efficiency. *J. exp. Educ.*, 1946, 15, 2-60.

204. LIPPITT, R. An experimental study of the effect of democratic and authoritarian group atmosphere. In *Studies in topological and vector psychology. I. Univer. Iowa Studies in Child Welfare*, 1940, 16. Pp. 45-65.

205. LITTLER, S. Causes of failure among elementary school teachers. *Sch. Home Educ.*, 1914, 33, 255-256.

206. LUDEMAN, W.W. What college freshmen think of their high school teachers. *Sch. Executives Magazine*, 1931, 50, 527-528.

207. LYNCH, J.M. Teacher rating trends psychologically examined. *Amer. Sch. Bd J.*, 1942, 104 (6), 27-28.

208. McAFEE, L.O. The reliability of the evidences of teaching efficiency secured in extension visitation. *Elem. Sch. J.*, 1930, 30, 746-754.

209. McCARTHA, C.W. The practice of teacher evaluation in the Southeast in 1948. *J. educ. Res.*, 1950, 44, 122-128.

210. McCOARD, W.B. Speech factors as related to teaching efficiency. *Speech Monogr.*, 1944, 11, 53-64.

211. MacDONALD, MARION E. Students' opinions as regards desirable and undesirable qualifications and practices of their teachers in teacher-training institutions. *Educ. Adm. Supervis.*, 1931, 17, 139-146.

212. McLAUGHLIN, J.O. A case study of teachers judged successful and nonsuccessful. In Stanford Univer., *Abstracts of dissertations . . . 1930-31*. Pp. 206-210.

213. MADSEN, I.N. The prediction of teaching success. *Educ. Adm. Supervis.*, 1927, 13, 39-47.

214. MAJOR, C.L. The percentile ranking on the Ohio State University psychological test as a factor in forecasting the success of teachers in training. *Sch. & Soc.*, 1938, 47, 582-584.

215. MALAN, C.T. What are the most desirable character traits of teachers? *Education*, 1931, 52, 220-226.

216. MANAHAN, J.L., and JARMAN, A.M. A comparison of superior and inferior teachers. *Amer. Sch. Bd J.*, 1935, 90 (4), 23-24.

217. MARKT, ANNA R., and GILLILAND, A.R. A critical analysis of the George Washington University Teaching Aptitude Test. *Educ. Adm. Supervis.*, 1929, 15, 660-666.

218. MARTIN, LYCIA O. *The prediction of success for students in teacher education.* New York: Teachers Coll., Columbia Univer., 1944.

219. MARTINDALE, F.E. Situational factors in teacher placement and success. *J. exp. Educ.*, 1951, 20, 121-177.

220. MATHEWS, L.H. An item analysis of measures of teaching ability. *J. educ. Res.*, 1940, 33, 576-580.

221. MEAD, A.R. Qualities of merit in good and poor teachers. *J. educ. Res.*, 1929, 20, 239-259.

222. MEAD, A.R., and HOLLEY, C.E. Forecasting success in practice teaching. *J. educ. Psychol.*, 1916, 7, 495-497.

223. MECHAM, G.P. A study of emotional instability of teachers and their pupils. *George Peabody Coll. Contr. Educ.*, 1941, No. 312.

224. MENON, T.K.N., and SUKLA, M.M. Intelligence and teaching ability. *Indian J. Psychol.*, 1946, 21, 33-38.

225. MERIAM, J.L. Normal school education and efficiency in teaching. *Teach. Coll. Contr. Educ.*, 1905, No. 1.

226. MESSIER, W.A. Are you the "best teacher." *Grade Teach.*, 1932, 49, 800-801; 829.

227. MICHAELIS, J.U., and TYLER, F.T. MMPI and student teaching. *J. appl. Psychol.*, 1951, 35, 122-124.

228. MOODY, F.E. The correlation of the professional training with the teaching success of normal-school graduates. *Sch. Rev.*, 1918, 26, 180-198.

229. MORGAN, A.L. Present status of teacher rating in the United States. Tex. Outlook, 1944, 28 (2), 21-22.

230. MORRIS, ELIZABETH H. Personal traits and success in teaching. Teach. Coll. Contr. Educ., 1929, No. 342.

231. MORRISON, R.H. Factors causing failure in teaching. J. educ. Res., 1927, 16, 98-105.

232. MORSH, J.E., and SWANSON, R.A. The relation of technical knowledge and other variables to ratings of instructor competence. San Antonio, Tex.: Human Resources Research Center, Lackland Air Force Base, April 1952. (Research Note Tech 52-1.)

233. MORTON, R.L. Qualities of merit in secondary teachers. Educ. Adm. Supervis., 1919, 5, 225-238.

234. MOSES, CLEDA V. Why high-school teachers fail. Sch. Home Educ., 1914, 33, 166-169.

235. MOSS, F.A., LOMAN, W., and HUNT, THELMA. Impersonal measurement of teaching. Educ. Rec., 1929, 10, 40-50.

236. MOTT, E.B. Teacher failures in the public schools. Agric. Educ., 1950, 22, 208; 210; 213.

237. NANNINGA, S.P. Teacher failures in high school. Sch. & Soc., 1924, 19, 79-82.

238. NANNINGA, S.P. Estimates of teachers in service made by graduate students as compared with estimates made by principal and assistant principal. Sch. Rev., 1928, 36, 622-626.

239. National Education Association. Teachers' salaries and salary trends in 1923. Nat. Educ. Res. Bull., 1923, 1, 1-115.

240. National Education Association. Practices affecting teacher personnel. Nat. Educ. Res. Bull., 1928, 6, 205-255.

241. National Education Association. Administrative practices affecting classroom teachers. Part II. The retention, promotion, and improvement of teachers. Nat.Educ. Res. Bull., 1932, 10, 33-76.

242. National Education Association. Teacher personnel procedures: Employment conditions in service. Nat. Educ. Res. Bull., 1942, 20, 81-116.

243. NEEL, MARY O., and MEAD, A.R. Correlations between certain group factors in preparation of secondary school teachers. Educ. Adm. Supervis., 1931, 17, 675-676.

244. NEMEC, LOIS G. Relationship between teacher certification and education in Wisconsin: A study of their effects on beginning teachers. J. exp. Educ., 1946, 15, 101-132.

245. NEWMARK, D. Students' opinions of their best and poorest teachers. Elem. Sch. J., 1929, 29, 576-585.

246. NUTTING, E.P. Does quantitative training produce better teaching? Sch. Executive, 1943, 62 (6), 21.

247. ODENWELLER, A.L. Predicting the quality of teaching, the predictive value of certain traits for effectiveness in teaching. Teach. Coll Contr. Educ., 1936, No. 676.

248. OLSON, W.C., and WILKINSON, MURIEL M. Teacher personality as revealed by the amount and kind of verbal direction used in behavior control. Educ. Adm. Supervis., 1938, 24, 81-93.

249. ORLEANS, J.S., CLARKE, D., OSTREICHER, L., and STANDLEE, L. Some preliminary thoughts on the criteria of teacher effectiveness. J. educ. Res., 1952, 45, 641-648.

250. OSBURN, W.J. The personal characteristics of the teacher. Educ. Adm. Supervis., 1920, 6, 74-85.

251. PAYNE, E.G. Scholarship and success in teaching. J. educ. Psychol., 1918, 9, 217-219.

252. PETERS, C.C., and VAN VOORHIS, W.R. Statistical procedures and their mathematical bases. New York: McGraw-Hill, 1940.

253. PETERS, D.W. The status of the married woman teacher. Teach. Coll. Contr. Educ., 1934, No. 603.

254. PETERSON, H.A., OBOURN, G., WALLACE, HAZEL, and SMITH, Q.W. Relation of scholarship during college career to success in teaching judged by salary. Educ. Adm. Supervis., 1934, 20, 625-628.

255. PETERSON, ODA K., and COOK, W.A. Score cards and rating sheets in teacher training. Educ. Method, 1930, 9, 322-330.

256. PHILLIPS, W.S. An analysis of certain characteristics of active and prospective teachers. George Peabody Coll. Contr. Educ., 1935, No. 161.

257. PORTER, W.A. Pupil evaluation of practice teaching. J. educ. Res., 1942, 35, 700-704.

258. POSEY, C.W. A new answer to an old problem--shall we rate teachers? Amer. Sch. Bd J., 1944, 108 (5), 34-35.

259. PRESTON, H.O. The development of a procedure for evaluating officers in the United States Air Force. Pittsburgh: American Institute for Research, 1948.

260. PYLE, W.H. Intelligence and teaching, an experimental study. Educ. Adm. Supervis., 1927, 13, 433-448.

261. PYLE, W.H. The relation between intelligence and teaching success: A supplementary study. Educ. Adm. Supervis., 1928, 14, 257-267.

262. REAVIS, W.C., and COOPER, D.H. Evaluation of teacher merit in city school systems. Suppl. educ. Monogr., 1945, No. 59.

263. REED, ANNA Y. The effective and the ineffective college teacher. New York: American Book, 1935.

264. REMMERS, H.H. The college professor as the student sees him. Purdue Univer. Stud. higher Educ., 1929, No. 11.

265. REMMERS, H.H. To what extent do grades influence student ratings of instructors? J. educ. Res., 1930, 21, 314-317.

266. REMMERS, H.H. Reliability and halo effect of high school and college students' judgments of their teachers. J. appl. Psychol., 1934, 18, 619-630.

267. REMMERS, H.H., and BRANDENBURG, G.C. Experimental data on the Purdue Rating Scale for Instructors. Educ. Adm. Supervis., 1927, 13, 519-527.

268. REMMERS, H.H., DAVENPORT, K.S., and POTTER, A.A. The best and the worst teachers of engineering. Purdue Univer. Stud. higher Educ., 1946, No. 57.

269. REMMERS, H.H., MARTIN, F.D., and ELLIOTT, D.N. Are students' ratings of instructors related to their grades? In H.H. Remmers (Ed.), Student achievement and instructor evaluation in chemistry. Purdue Univer. Stud. higher Educ., 1949, No. 66. Pp. 17-26.

270. REMMERS, H.H., SHOCK, N.W., and KELLY, E.L. An empirical study of the validity of the Spearman-Brown formula as applied to the Purdue Rating Scale. J. educ. Psychol., 1927, 18, 187-195.

271. RETAN, G.A. Emotional instability and teaching success. J. educ. Res., 1943, 37, 135-141.

272. RICHARDSON, M.W. Forced-choice performance reports: A modern merit-rating method. Personnel, 1949, 26, 205-212.

273. RICHEY, H.W., and BERKSHIRE, J.R. *Job satisfaction of Air Force technical school instructors.* San Antonio, Tex.: Human Resources Research Center, Lackland Air Force Base, November 1952. (*Research Note* Tech 52-12.)

274. RICHEY, R.W., and FOX, W.H. A study of high school students with regard to teachers and teaching. *Ind. Univer. Sch. Educ. Bull.*, 1951, 27, 9-63.

275. RIESCH, K.P. A study of some factors in pupil growth. *J. exp. Educ.*, 1949, 18, 31-55.

276. RILEY, J.W., Jr., RYAN, B.F., and LIFSHITZ, MARCIA. *The student looks at his teacher.* New Brunswick, N.J.: Rutgers Univer. Press, 1950.

277. RINGNESS, T.A. Relationships between certain attitudes towards teaching and teaching success. *J. exp. Educ.*, 1952, 21, 1-55.

278. RITTER, E.L. Rating of teachers in Indiana. *Elem. Sch. J.*, 1918, 18, 740-756.

279. ROBERTS, A.C., and DRAPER, E.M. *The high-school principal as administrator, supervisor, and director of extracurricular activities.* New York: Heath, 1927.

280. ROLFE, J.F. The measurement of teaching ability. Study number two. *J. exp. Educ.*, 1945, 14, 52-74.

281. ROOT, A.R. Student ratings of teachers. *J. higher Educ.*, 1931, 2, 311-315.

282. ROSTKER, L.E. The measurement of teaching ability. Study number one. *J. exp. Educ.*, 1945, 14, 6-51.

283. RUEDIGER, W.C., and STRAYER, G.D. The qualities of merit in teachers. *J. educ. Psychol.*, 1910, 1, 272-278.

284. RUGG, H. Is the rating of human character practicable? *J. educ. Psychol.*, 1921, 12, 425-438; 485-501; 1922, 13, 81-93.

285. RYANS, D.G. The results of internal consistency and external validation procedures applied in the analysis of test items measuring professional information. *Educ. psychol. Measmt*, 1951, 11, 549-560.

286. RYANS, D.G. A study of the extent of association of certain professional and personal data with judged effectiveness of teacher behavior. *J. exp. Educ.*, 1951, 20, 67-77.

287. RYANS, D.G. A study of criterion data (a factor analysis of teacher behaviors in the elementary school). *Educ. psychol. Measmt*, 1952, 12, 333-344.

288. RYANS, D.G., and WANDT, E. Investigations of personal and social characteristics of teachers. *J. Teach. Educ.*, 1952, 3, 228-231.

289. RYANS, D.G., and WANDT, E. A factor analysis of observed teacher behaviors in the secondary school: A study of criterion data. *Educ. psychol. Measmt*, 1952, 12, 574-586.

290. RYLE, FLORENCE. Qualities that students admire in teachers. *Calif. Quart. secondary Educ.*, 1928, 4, 82-85.

291. SAMUELSON, E.E. The evaluation of teachers and teaching. *Sch. Executive*, 1941, 60 (9), 15-16; 27.

292. SANDIFORD, P., CAMERON, M.A., CONWAY, C.B., and LONG, J.A. Forecasting teaching ability. *Univer. Toronto educ. Res. Bull.*, 1937, No. 8.

293. SCHAFFIE, A.E.F. The pupil checks the teacher. *Sch. Executives Magazine*, 1931, 51, 151-153.

294. SCHELLHAMMER, F.M. Rating the practice teacher. *Sch. Executive*, 1940, 60, (2), 32-33.

295. SCHMID, J., Jr. Factor analyses of prospective teachers' differences. *J. exp. Educ.*, 1950, 18, 287-319.

296. SCHUTTE, T.H. The teacher, through the students' eyes. *Amer. Sch. Bd J.*, 1926, 73 (3), 72-79.

297. SCHWARTZ, A.N. A study of the discriminating efficiency of certain tests of the primary source personality traits of teachers. *J. exp. Educ.*, 1950, 19, 63-93.

298. SEAGOE, MAY V. Prognostic tests and teaching success. *J. educ. Res.*, 1945, 38, 685-690.

299. SEAGOE, MAY V. Prediction of in-service success in teaching. *J. educ. Res.*, 1946, 39, 658-663.

300. SEELEY, L.C. *Preliminary validation of the instructors evaluation report.* Washington: Office of Naval Research, April 1949. (*Technical Report*, SDC 383-1-9.)

301. SENGER, H.L. Status teacher-rating in 1936. *Amer. Sch. Bd J.*, 1936, 93 (3), 56.

302. SEYFERT, W.C., and TYNDAL, B.S. An evaluation of differences in teaching ability. *J. educ. Res.*, 1934, 28, 10-15.

303. SHANNON, J.R. *Personal and social traits requisite for high grade teaching in secondary school*. Terre Haute, Ind.: State Normal Press, 1928.

304. SHANNON, J.R. A comparison of three means for measuring efficiency in teaching. *J. educ. Res.*, 1936, 29, 501-508.

305. SHANNON, J.R. A comparison of highly successful teachers, failing teachers, and average teachers at the time of their graduation from Indiana State Teachers College. *Educ. Adm. Supervis.*, 1940, 26, 43-51.

306. SHANNON, J.R. Elements of excellence in teaching. *Educ. Adm. Supervis.*, 1941, 27, 168-176.

307. SHANNON, J.R. A measure of the validity of attention scores. *J. educ. Res.*, 1942, 35, 623-631.

308. SHIELS, A. The rating of teachers in the New York City public schools. *Sch. & Soc.*, 1915, 2, 752-754.

309. SHULTZ, I.T. *A descriptive and predictive study of a class in a school of education*. Unpublished doctor's dissertation, Univer. of Pennsylvania, 1928.

310. SIMMONS, EDNA. *Correlation of administrative ratings of teachers and pupil achievement*. Unpublished doctor's dissertation, George Peabody College for Teachers, 1932.

311. SIMON, D.L. Personal reasons for the dismissal of teachers in smaller schools. *J. educ. Res.*, 1936, 29, 585-588.

312. SIMPSON, R.H., GAIER, E.L., and JONES, S. A study of resourcefulness in attacking professional problems. *Sch. Rev.*, 1952, 60, 535-540.

313. SISSON, E.D. Forced-choice--the new Army rating. *Personnel Psychol.*, 1948, 1, 365-381.

314. SMALZRIED, N.T., and REMMERS, H.H. A factor analysis of the Purdue Rating Scale for Instructors. *J. educ. Psychol.*, 1943, 34, 363-367.

315. SMELTZER, C.H., and HARTER, R.S. Comparison of anonymous and signed ratings of teachers. *Educ. Outlook*, 1934, 8, 76-84.

316. SMITH, A.A. What is good college teaching? *J. higher Educ.*, 1944, 15, 216-218.

317. SMITH, A.A. What traits do high school pupils admire in teachers? *High Sch. J.*, 1945, 28, 279-286.

318. SMITH, E.L. A critical analysis of rating sheets now in use for rating student-teachers. *Educ. Adm. Supervis.*, 1936, 22, 179-189.

319. SODERQUIST, H.O. Participation in extracurricular activities in high school or college and subsequent success in teaching adults. *Sch. & Soc.*, 1935, 42, 607-609.

320. SOMERS, G.T. Pedagogical prognosis. Predicting the success of prospective teachers. *Teach. Coll. Contr. Educ.*, 1923, No. 140.

321. STALNAKER, J.M., and REMMERS, H.H. Can students discriminate traits associated with success in teaching? *J. appl. Psychol.*, 1928, 12, 602-610.

322. STARRAK, J.A. Student rating of instruction. *J. higher Educ.*, 1934, 5, 88-90.

323. STEPHENS, J.M., and LICHTENSTEIN, A. Factors associated with success in teaching grade five arithmetic. *J. educ. Res.*, 1947, 40, 683-694.

324. STEWART, MAY L. A study of success and failure in eleven years of rural teacher training. *Educ. Adm. Supervis.*, 1940, 26, 372-378.

325. STOLUROW, L.M., IRION, A.L., and PASCAL, G.R. The selection and training of gunnery instructors. In N. Hobbs (Ed.), *Psychological research on flexible gunnery training*. AAF Aviation Psychology Program Research Report No. 11, 1947, Pp. 337-381.

326. STUIT, D.B. Scholarship as a factor in teaching success. *Sch. & Soc.*, 1937, 46, 382-384.

327. STUIT, D.B., and EBEL, R.L. Instructor rating at a large state university. *Coll. & Univer.*, 1952, 27, 247-254.

328. STUMP, N.F. A comparative study of two teaching aptitude tests. *J. educ. Psychol.*, 1937, 28, 595-600.

329. TAYLOR, E.K., and WHERRY, R.J. A study of leniency in two rating systems. *Personnel Psychol.*, 1951, 4, 39-47.

330. TAYLOR, H.R. The influence of the teacher on relative class standing in arithmetic fundamentals and reading comprehension. *Nat. Soc. Stud. Educ. Yearb.* 1928, 27 (Part II), 97-110.

331. TAYLOR, H.R. Teacher influence on class achievement: A study of the relationship of estimated teaching ability to pupil achievement in reading and arithmetic. *Genet. Psychol. Monogr.*, 1930, 7, 81-175.

332. THURSTONE, L.L. The method of paired comparisons for social values. *J. abnorm. soc. Psychol.*, 1927, 21, 384-400.

333. TIEGS, E.W. *An evaluation of some techniques of teacher selection.* Bloomington, Ill.: Public School Publishing Co., 1928.

334. TOSTLEBE, M.F. Analysis of the relative importance of the success factors common in the training of teachers for the one-room rural school. *J. educ. Res.*, 1937, 30, 397-402.

335. TRAXLER, A.E. Are students in teachers colleges greatly inferior in ability? *Sch. & Soc.*, 1946, 63, 105-107.

336. TUDHOPE, W.B. A study of the training college final teaching mark as a criterion of future success in the teaching profession. *Brit. J. educ. Psychol.*, 1942, 12, 167-171; 1943, 13, 16-23.

337. TYLER, F.T. Personality tests and teaching ability. *Canad. J. Psychol.*, 1949, 3, 30-37.

338. TYLER, LEONA E. *The psychology of human differences.* New York: Appleton-Century-Crofts, Inc., 1947.

339. ULLMAN, R.R. The prediction of teaching success. *Educ. Adm. Supervis.*, 1930, 16, 598-608.

340. ULLMAN, R.R. *The prognostic value of certain factors related to teaching success.* Ashland, Ohio: A.L. Garber, 1931.

341. U. S. Office of Indian Affairs. A rating scale for Indian service teachers. *Progr. Educ.*, 1940, 17, 363-366.

342. UPSHALL, C.C. A ten-year study of two groups of teachers college students of contrasting ability. *J. Amer. Ass. Collegiate Registrars*, 1942, 18, 36-44.

343. UPSHALL, C.C. The validity of composite faculty judgment as a method of identifying undesirable prospective elementary school teachers. *J. educ. Res.*, 1942, 35, 694-699.

344. VON HADEN, H.I. An evaluation of certain types of personal data employed in the prediction of teaching efficiency. *J. Exp. Educ.*, 1946, 15, 61-84.

345. WADDELL, C.W. The prognostic value of Army Alpha scores for success in practice-teaching. *Educ. Adm. Supervis.*, 1927, 13, 577-592.

346. WAGENHORST, L.H. The relation between ratings of student teachers in college and success in first year of teaching. Educ. Adm. Supervis., 1930, 16, 249-253.

347. WARD, L.B., and KIRK, S.A. Studies in the selection of students for a teachers college. J. educ. Res., 1942, 35, 665-672.

348. WARD, W.D., REMMERS, H.H., and SCHMALZRIED, N.T. The training of teaching-personality by means of student-ratings. Sch. & Soc., 1941, 53, 189-192.

349. WELLS, F.L. A statistical study of literary merit. Arch. Psychol., 1907, No. 7.

350. WHERRY, R.J. Comparative validity of the WD AGO Form No. 67 and the FCL-2 according to various breakdowns., Washington: Personnel Research Section, Adjutant General's Office, December 1945. (Report No. 671.)

351. WHITNEY, F.L. The intelligence, preparation, and teaching skill of state normal school graduates in the United States., Minneapolis: Univer. of Minnesota, 1922.

352. WHITNEY, F.L. The prediction of teaching success. J. Educ. Res. Monogr., 1924, No. 6.

353. WHITNEY, F.L., and FRASIER, C.M. The relation of intelligence to student teaching success. Peabody J. Educ., 1930, 8, 3-6.

354. WILSON, W.R. Students rating teachers. J. higher Educ., 1932, 3, 75-82.

355. WITHALL, J. The development of a technique for the measurement of social-emotional climate in classrooms. J. exp. Educ., 1949, 17, 347-361.

356. WITTY, P.A. An analysis of the personality traits of the effective teacher. J. educ. Res., 1947, 40, 662-671.

357. WITTY, P.A. Evaluation of studies of the characteristics of the effective teacher. In Improving educational research; Official Report Amer. educ. res. Ass., 1948. Pp. 198-204.

358. WOELLNER, R.C. Evaluation of apprentice teachers. Sch. Rev., 1941, 49, 267-271.

359. WRIGHTSTONE, J.W. Measuring teacher conduct of class discussion. Elem. Sch. J., 1934, 34, 454-460.

360. WRIGHTSTONE, J.W. Constructing an observational technique. *Teach. Coll. Rec.*, 1935, 37, 1-9.

361. WRIGHTSTONE, J.W. Rating methods. In W.S. Monroe (Ed.), *Encyclopedia of educational research*. New York: Macmillan, 1950. Pp. 961-964.

362. YOUNG, F. Efficiency of high school teachers as measured by principals' ratings. *Tex. Outlook*, 1937, 21 (1), 24-25.

363. YOUNG, F. Some factors affecting teaching efficiency. *J. educ. Res.*, 1939, 32, 649-652.

364. ZANT, J.H. Predicting success in practice-teaching. *Educ. Adm. Supervis.*, 1928, 14, 664-670.

Reviews and Bibliographies

365. ANDERSON, E.W. Techniques of research used in the field of teacher personnel. *Rev. educ. Res.*, 1934, 4, 15-20.

366. BARR, A.S. Measurement of teaching ability. *Rev. educ. Res.*, 1940, 10, 182-184; 267-268.

367. BARR, A.S. The measurement and prediction of teaching efficiency. *Rev. educ. Res.*, 1943, 13, 218-223.

368. BARR, A.S. The measurement and prediction of teaching efficiency. *Rev. educ. Res.*, 1946, 16, 203-208.

369. BARR, A.S. The measurement and prediction of teaching efficiency: A summary of investigations. *J. exp. Educ.*, 1948, 16, 203-283.

370. BARR, A.S. Measurement and prediction of teaching success. *Rev. educ. Res.*, 1949, 19, 185-190.

371. BARR, A.S. The measurement of teacher characteristics and prediction of teaching efficiency. *Rev. educ. Res.*, 1952, 22, 169-174.

372. BARR, A.S., and DOUGLAS, LOIS. The pre-training selection of teachers. *J. educ. Res.*, 1934, 28, 92-117.

373. BEECHER, D.E. *Evaluation of teaching backgrounds and concepts*. Syracuse Univer. Press, 1949.

374. BETTS, G.L. The education of teachers evaluated through measurement of teaching ability. In *National survey of the education of teachers. U. S. Off. Educ. Bull.*, 1933, No. 10, (5), 87-153.

375. BUTSCH, R.L. Teacher rating. *Rev. educ. Res.*, 1931, 1, 99-107; 149-152.

376. BYERS, B.H. Speech in the prediction of teaching success. *Peabody J. Educ.*, 1950, 28, 80-87.

377. COLE, LUELLA. *The background for college teaching.* New York: Farrar and Rinehart, 1940.

378. COREY, S.M. The present state of ignorance about the factors affecting teacher success. *Educ. Adm. Supervis.*, 1932, 18, 481-490.

379. COREY, S.M. What are the factors involved in the success of high school teachers? *North Central Ass. Quart.*, 1935, 10, 224-231.

380. DOMAS, S.J., and TIEDEMAN, D.V. Teacher competence: An annotated bibliography. *J. exp. Educ.*, 1950, 19, 101-218.

381. DURFLINGER, G.W. A study of recent findings on the prediction of teaching success. *Educ. Adm. Supervis.*, 1948, 34, 321-336.

382. EVANS, KATHLEEN M. A critical survey of methods of assessing teaching ability. *Brit. J. educ. Psychol.*, 1951, 21, 89-95.

383. FRAZIER, B.W. Education of teachers: Selected bibliography, October 1, 1935 to January 1, 1941. *U. S. Off. Educ. Bull.*, 1941, No. 2.

384. JEWETT, IDA A. Summary of studies on teacher selection. In M. M. Stroh, I.A. Jewett, and V.M. Butler. *Better selection of better teachers.* Washington: Delta Kappa Gamma Soc., 1943. Pp. 82-91.

385. LANCASTER, J.H. A guide to the literature on the education of teachers. *Educ. Adm. Supervis.*, 1933, 19, 363-372.

386. MAHLER, W.R. *Twenty years of merit rating, 1926-1946.* New York: The Psychological Corp., 1947.

387. MONROE, W.S. Controlled experimentation as a means of evaluating methods of teaching. *Rev. educ. Res.*, 1934, 4, 36-42.

388. TORGERSON, T.L. The measurement and prediction of teaching ability. *Rev. educ. Res.*, 1934, 4, 261-266; 329-330.

389. TORGERSON, T.L. Measurement and prediction of teaching ability. *Rev educ. Res.*, 1937, 7, 242-246; 319-320.

390. TROW, W.C., and McLOUTH, FLORENCE. An improvement card for student-teachers. *Educ. Adm. Supervis.*, 1929, 15, 1-10; 127-133.

391. U. S. Office of Education. Bibliography of research studies in education. U. S. Off. Educ. Bull., Nos. 22, 1928; 36, 1929; 23, 1930; 13, 1931; 16, 1932; 6, 1933; 7, 1934; 5, 1935; 5, 1936; 6, 1937; 5, 1938; 5, 1939; 5, 1940; 5, 1941.

392. YAUKEY, J.V., and ANDERSON, P.L. A review of the literature on the factors conditioning teaching success. Educ. Adm. Supervis., 1933, 19, 511-520.

Manuscript received 23 November 1953.